DOCUMENT RESUME

ED 077 537 LI 004 364

TITLE	QUIS: Queen's University Information Systems; Report of Activities for the Period Ending on 31st December, 1972.
INSTITUTION	Queen's Univ., Belfast (Northern Ireland). Dept. of Computer Science.
SPONS AGENCY	Office for Scientific and Technical Information, London (England).
PUB DATE	Jan 73
NOTE	146p.; (38 References)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Annual Reports; Bibliographic Citations; Data Bases; Information Dissemination; *Information Retrieval; *Information Systems; *On Line Systems; *Physics

ABSTRACT

Two computer information systems, one for reference retrieval and the other for the retrieval of numerical data on atomic and molecular physics are being developed at the Queen's University of Belfast. The reference retrieval system is based on the atomic and molecular physics section of the INSPEC tapes. The tapes contain complete abstracts of the documents which are incorporated into the system, and from these a thesaurus of relevant keywords is constructed. The data base is searched using these keywords. There are over 7000 documents indexed and the thesaurus contains about 3000 terms. The numerical data system allows a user to retrieve and manipulate with numerical data. At present, the data base consists of atom-atom potentials which are extracted from the literature. For each state is stored the available potential data along with a reference to the source, an estimate of the accuracy and the range of validity. The potentials can take various forms, for example, a table of values or the parameters of a formula of known form. At present six forms of potential curve fits can be accommodated. The fits are stored in atomic units but can be outputted in the units chosen by the user. The system allows the user to manipulate with the stored data. (Author/SJ)

ED 077537

FILMED FROM BEST AVAILABLE COPY

The Queen's University of Belfast

Department of Computer Science

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

QUB/OSTI

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

QUIS

QUEEN'S UNIVERSITY INFORMATION SYSTEMS

PART A

Project QUOBIRD

Queen's University On-line Bibliographic Information Retrieval and Dissemination System

PART B

Project QUODAMP

Queen's University Databank on Atomic and Molecular Physics

Report of activities for the period ending on
31st December, 1972

January, 1973.

LI 004 364

ACKNOWLEDGMENTS

The research reported in this document was supported by a grant from the Office of Scientific and Technical Information, Department of Education and Science and by Queen's University, Belfast

PERSONNEL

Director:	Dr. F. J. Smith
Manager:	Mr. L. D. Higgins
Research and associated Staff:	Dr. H. O'Hara, Systems analyst and project leader Dr. J. Boyle, Systems analyst and information scientist (until 30th September, 1972) Mr. R. McDonough, Systems analyst and information scientist (from 1st October, 1972) Mr. R. Gregg, Senior programmer Mrs. J. Stewart, Senior library assistant Mr. J. McClean, Programmer Mr. G. McGlinchey, Programmer Miss M. Bell, Control clerk Mrs. K. Moyna, Secretary (part-time)
Research students:	Mr. J. Jamieson Mr. F. Jabbour Mr. J. Magaud Mrs. M. McCloy
Academic associates:	Professor C.A.R. Hoare, Professor in Computer Science Dr. A.E. Kingston, Reader in Applied Mathematics Dr. D.C.S. Allison, Senior lecturer in Computer Science Mr. P. Quigg, Lecturer in Library School Mr. R. Kimber, Lecturer in Library School Mr. H. McKeown, Lecturer in St.Joseph's Training College
Supporting Systems Group: (Computer Centre)	Mr. R.A. McLaughlin, Manager of Systems Development Mr. J. Warne, Systems consultant Mr. C. Johnson, Senior programmer Mr. W. Carson, Programmer Dr. K. Devine, Numerical analyst Mrs. M. Carville, Senior programmer Mr. T. Wright, Senior programmer

PART A

Project QUOBIRD

CONTENTS

		Page
Publications and Reports		
1.	INTRODUCTION	1
1.1	Background	1
1.2	Position at 1/1/72	4
1.3	Position at 1/1/73	5
1.4	Delays	7
1.5	Conferences and Demonstrations	8
2.	PROPOSALS ACCEPTED BY OSTI	10
2.1	Continuation of pilot project	10
2.2	Query formulation	13
2.3	Reasons for OSTI support	14
3.	PROBLEMS WITH EXPANSION	15
3.1	Off-line indexing	15
3.2	Prior dictionary	16
3.3	Synonyms	16
3.4	Housekeeping	17
3.5	File security problems	19
4.	COMMAND LANGUAGE IMPROVEMENTS	22
4.1	Introduction	22
4.2	Description of on-line retrieval system from physics abstracts	22
4.3	Applied Mathematics, Physics and Computer Science library	24
4.4	Hyphenated words	25
4.5	Suppression of superfluous printing during retrieval	27
4.6	Off-line printout	27
5.	ADDITIONAL RESEARCH	29
5.1	Fortran or Assembly language	29
5.2	Efficiency of hash indexing	30
5.3	Machine-independence of data bases and the programs which manipulate them	32
5.4	Data compression	33
6.	ASSESSMENT OF PRESENT POSITION	35
References		39
Appendices		

PUBLICATIONS AND REPORTS

1. "On-line Subject Indexing and Retrieval" by L.D.Higgins and F.J. Smith. Program, Vol.3, 147-56, 1969.

2. "A Study of an Interactive Retrieval System" by L.D. Higgins. Thesis, Queen's University, Belfast, 1969.

3. "A Study of an Interactive Storage and Retrieval System" by J. Ashraf. Thesis, Queen's University, Belfast, 1970.

4. "A Comparative Study of Two Information Retrieval Systems" by P.M. MacLaverty. Thesis, Queen's University, Belfast, 1970.

5. "Queen's University On-line Bibliographic Information Retrieval and Dissemination System" Annual Report to D.P.C., 1970.

6. "A Study of Re-entrant Systems" by W.A.V. Pringle. Thesis, Queen's University, Belfast, 1971.

7. Efficient On-line Document Retrieval" by F.J.Smith and L.D. Higgins. Datafair '71. Abstracts, 1971.

8. "Interactive Reference Retrieval in Large Files" by M. Carville, L.D. Higgins and F.J. Smith. Paper read at Cranfield International Conference on Mechanised Information Storage and Retrieval Systems, 1971. Infor. Stor. Retr., Vol. 7, 205-10, 1971.

9. "Disc Access Algorithms" by L.D. Higgins and F.J. Smith. Comp. Journal, Vol. 14, No. 3, 249-53, 1971.

10, "Queen's University On-line Bibliographic Information Retrieval and Dissemination System" Annual Activity Report for the period ending 31st December, 1971.

11. "On-line Data Banks in Atomic and Molecular Physics" by F.J. Smith. Paper read at the VII International Conference on the Physics of Electronic and Atomic Collisions, VII IPEAC, 1971 - North-Holland, 1972.

12. "Short Feasibility Study on using the Post Office Telex Network for On-line Computer Information Retrieval" Special Report, SR3, January 1972.

13. "The Data Base Access Problem" Special Report, SR5, May 1972.

14. "Alphabetic Printout of the Dictionary as exists in our Inspec disc file at present" Special Report, SR6, June 1972.

15. "QUIS Projects" by L.D. Higgins. Paper read at NATO Advanced Study Institute Course on On-line Mechanised Information Retrieval Systems at Lyngby, July/August, 1972.

16. "The BIRD On-line Retrieval System" by L.D. Higgins. Paper read at the Universities' Computer Science Colloquium at Edinburgh, September, 1972.

1. INTRODUCTION

1.1 Background

In our studies of on-line information retrieval systems over the last five years, we have constantly had in mind the situation which we expect will prevail at the end of this decade when most scientists, engineers, doctors, lawyers, etc. have beside their desks simple and flexible computer terminals linked to a local computer and through this local computer to a data and computer network beyond. They will use the computer terminal everyday for either calculations or information, or both. It may be that most of the time the information they store and retrieve will be in a small personalised data bank on their local computer, but they will also be able to interrogate other data banks using a special data network or using the telephone network. We take the view that the cost of computing, data storage and transmission combined can be lower than the cost of maintaining the data, provided that the software controlling the information systems is efficient and well designed. Because the computing and transmission costs will be low and because the biggest cost will be the maintaining of the quality of the data we expect that there will be several data banks in different parts of Europe and the world specialising in particular areas. For this reason we chose in Belfast to carry out our research with data in the field of Atomic and Molecular Physics, because Queen's University, Belfast is a leading world centre of research in this

discipline. Thus we can draw on the expertise available in Belfast and when necessary employ the experienced staff we need for maintaining the data.

We expect that our numerical data system (which is discussed in Part B of this report) will be the only one of its type in the world. We aim to make it so cost-effective that in spite of the telephone or transmission costs, the system will be worth interrogating from California or Tokyo. It will be so specific and so difficult to keep up to date that it will probably not be worth keeping copies of it in other locations (e.g. U.S.A.). On the other hand, we expect that our reference system will be accessed frequently enough to make it expedient to have copies available at 3 or 4 places in the world (Colorado? Tashkent? Tokyo?). These other copies will be updated automatically every week from Belfast. The European data system will be interrogated irregularly by the workers in atomic and molecular physics throughout Europe and by others in different fields, particularly chemists, engineers and nuclear physicists requiring occasional information on atomic and molecular physics. Many users will only use the system once a year or less; few users will use it more than 5 or 6 times in a year. The searches will normally be retroactive for a specified number of years. The data file will contain almost the whole literature in the narrow field of specialisation. Complete data files on other subjects will be kept at various places and will be within the reach of any scientist who can afford to pay the $2 or $3 fee. (We would hope that to scientists the data would

be free and the system financed by an international body. At a cost of about $500,000 per annum all workers in atomic and molecular physics throughout the world could receive a free service.)

Because the system would be used infrequently by each scientist and because he cannot be expected to learn and remember a complex system, using it only 3 or 4 times a year, the retrieval system will have to be very simple. This excludes the use of complex Boolean statements and necessitates some form of question/answer dialogue based on simple multiple choice questions in natural language, the user having to initiate as few responses as possible. It also implies a system which is self-instructive and which can be used without previous knowledge or experience by any intelligent user.

It is apparent that a fairly large file of references would be needed and a rapid response required, not just because some long distance telephone calls would be involved but also because a large part of the cost of the retrieval would be proportionate to the length of the average response (other factors being equal). We found some five years ago that it was relatively easy to write an on-line system for a small number of documents - we were able to demonstrate one for 100 documents after only 6 months inexperienced programming - but it is a great deal more difficult to write a system which gives a realistic response and cost for a large number of references. It is also relatively easy to write a system using a controlled vocabulary and more difficult to use a free language thesaurus because of

its greater size, the different forms of words, problems with prefixes, synonyms, etc. However, we feel that, looking to the future, the free vocabulary more likely meets the needs of the user and it allows automatic indexing using significant words in an abstract or in part (or later all) of the text. The cost of professional indexers using a controlled thesaurus to index a paper is high and the advantages doubtful because significant terms can be extracted adequately and much more cheaply by the computer.

All of the factors outlined in the above paragraphs have influenced us to build and rebuild a system with a free vocabulary, references indexed by the significant words in the title and abstract, and the emphasis, up to the present, put on producing extremely efficient software to give the fastest response and lowest cost possible to the user. We have no doubt that once we have mastered the problem of producing efficient, flexible and modular software which is proven to deal at low cost with a large number of references, that it will be possible quickly and easily to adapt this software to produce many forms of retrieval systems and different query languages.

1.2 Position at 1/1/72

By the end of 1971 we had successfully built and demonstrated a pilot on-line system for the retrieval of journal references, taken from current INSPEC tapes. We had used this pilot system to study and develop efficient computing techniques for interactive retrieval work. The

pilot project had matured to the point where we could automatically retrieve information from a small file of approximately 1000 augmented catalogue records which included titles and abstracts. In parallel, a secondary database consisting of titles and chapter headings as well as bibliographic details of books had been implemented for a small departmental computer science library, allowing the user to interrogate and retrieve on-line the stored information. Detailed descriptions of the above work can be found in previous annual reports.

As a result of two years experience with the project there were a number of enhancements and improvements which suggested themselves, some of which we have implemented in the past year. These are described later in this report.

1.3 Position at 1/1/73

The work of adding new records to the data bases of atomic and molecular physics abstracts and books in the applied mathematics, physics and computer science libraries has continued throughout the year. Our main effort has been devoted to building up the abstracts file which over the year has increased from 1000 to over 6000 references. The references are extracted from the Physics Abstracts section on magnetic tapes prepared by Inspec. The file covers all papers published in the field over the previous two years and is now large enough to be useful. In increasing the data base sixfold, we were presented with many problems not expected and not encountered in the small

test data file. We describe these at a later stage in the report.

A new library, called the School of Applied Mathematics and Physics Library, which merges the books from the Applied Mathematics, Physics and Computer Science departmental libraries, has been set up as a branch of the main library at Queen's University. Almost all of the 1,000 books in this collection have been set up as a secondary data base which allows the retrieval of information from the titles and chapter headings of books. As new books are acquired by this library, their bibliographic details are coded in the appropriate format and they are added to the file.

The updating process of the Inspec records has been substantially changed from last year. Instead of updating entirely in an on-line mode an off-line version is first applied on all known words to the system leaving only a small number of words (new words) by comparison to be handled on-line. This has speeded up our work considerably and is discussed later.

The on-line control system, MCS (Multiplexor Control System), an operating system developed at Queen's University, Belfast, has been modified to allow more than one program to access the same data file simultaneously and also to allow more than one user to access the same program file at the same time. Each user is given a separate copy of the same retrieval program.

After a very successful demonstration of the retrieval system to the School of Applied Mathematics and Physics library committee, showing on-line retrieval both from the physics abstracts file and the merged library books, a

Visual Display Unit has been installed in the library to facilitate readers who may wish to use the system. As a result, we think we are now in a much better position to improve the user interface side of the system.

1.4 Delays

During the year our work has not been held up significantly, as might be expected by people living outside Belfast, by the constant sound of bombs and bullets, some very close indeed! Only a few days work were lost because of the "troubles". However, our work was held back considerably by the poor performance of the hardware, particularly the large Fixed Disc at the University's Computer Centre on which our work depended almost totally. Time after time files which had been built up and edited painstakingly were lost and the work had to be repeated; time after time software errors were traced after days of wasted effort to file corruptions which in theory should not have been possible if the hardware checking mechanisms had been working properly. Overall, we estimate that because of these difficulties the project was over one month behind schedule at the end of the year. These problems will continue in 1973; but fortunately the offending dev'ce, a Bryant Fixed Disc, is being replaced at the University Computing Centre by an ICL exchangeable disc, EDS 60. This should be more reliable.

1.5 Conferences and Demonstrations

During the summer of 1972, members of the information systems group attended and participated in a variety of conferences and meetings. Papers on OSTI supported research projects in Belfast were given at the NATO Advanced Study Institute in Denmark and the Universities' Computer Science Colloquium in Scotland. Copies of these papers and reports on the conferences are given in the appendices. In addition, on-line mechanised information storage and retrieval demonstrations, linked to the ICL 1907 at Queen's, were successfully displayed in Denmark and Scotland. The Codata Conference in France was attended by two members of the group and the On-line '72 conference in England was attended by one member of the group. Reports of these are also given in the appendices.

In particular, the demonstration in Scotland created a lot more interest than anticipated since, unlike the NATO one in Denmark, the colloquium held was of a general software nature, with information retrieval systems forming only a small part of the agenda. However, the talk stimulated such interest in the audience that a request was made to see a live demonstration of the system. This was arranged on the evening of the same day and for two hours three of the QUIS group present were busy answering queries of interest to an audience of about thirty people who were given the opportunity of trying out the system for themselves. The result was not only pleasing and satisfactory for the participants but was encouraging and reassuring for the QUIS research members at Queen's.

Other University departments to link into the system included the physics department at Stirling University, the mathematics department at Southampton University and the main library at Birmingham University. In the university here in Belfast we have given several demonstrations - among these was one given to the Engineering Department. A member of the group gave a talk and live demonstration at the New Polytechnic in Belfast, both of which were recorded.

The impact that the information retrieval team has created at Queen's has been such that a full term's course on Information Retrieval is now included in the syllabus for postgraduate studies for further degrees. The course is being given by the director of the group and will mainly deal with the work encountered in implementing our systems.

2. PROPOSALS ACCEPTED BY OSTI

Prior to the ending of the previous grant, proposals were submitted by us for a continuation of the work involved for a further two years from the beginning of 1972. A summary is given in this section of the parts of this proposal which OSTI agreed to support.

2.1 Continuation of Pilot Project

By the end of our previous OSTI grant period in December 1971, we had already spent some months in building up a small file of titles and abstracts on Atomic and Molecular Physics from the Inspec tapes, the size of the pilot file was then approximately 1000 records. The system had been designed to index and retrieve automatically from a data base of 1,000 to 10,000 references. By the end of the proposed two year period of the present grant it is expected that the data base will grow to around 10,000 references. In the process of this expansion a number of modifications and improvements are to be added to the system. These include the following:

(a) Off-line Indexing

An alternative version of the indexing program which will work mainly off-line is to be set up. This will work in a similar manner to the indexing program used to create the already existing small file. It will index each word in an abstract automatically if the word has arisen previously, but new words, instead of being displayed on the tele-typewriter, will be stored on a temporary file or backing store. They will later be printed out on the line-printer

with either a reference to, or the full text of, the abstract from which they come. The indexer will then process these words and using a housekeeping program to retrieve them from their temporary store he can add them to the thesaurus and update their associated entry lists, etc. on-line. This version of the indexing program will have the advantage that it will use mainly cheaper off-line overnight runs and the indexing time will not be wasted sitting at the teletypewriter while large portions of records are indexed automatically. The program can also be adapted to give a printout, when desired, of each abstract and a list of all the terms arising from that abstract. The facility will be useful for keeping periodic checks on the indexing and stemming of words. The original indexing program was written for on-line use, not because we thought it was better to do it that way, but because it gave us our first experience of on-line programming. We believe that off-line indexing, as we now propose, will be cheaper and result in better indexing.

(b) Prior Dictionary

In many subject fields people have already put considerable effort into building up thesauri. We feel that we can utilise some of this effort by providing a facility for incorporating a particular dictionary, or part of a dictionary, into any of our files before or after we start indexing any documents. To date the whole thesaurus is built from terms as they occur in the references which are indexed; in future it will be from both.

(c) Synonyms

So far our methods of dealing with synonyms have been simple and designed primarily to be only a means of dealing with different forms of the one word - e.g., electron, electrons, electronic. In the new system the user can, having presented a word to the system, be presented with a list of related words, all of which he can decide to use in his search, depending on the context of his query. The difficulty is the number of questions we must put to the user to get this information from him, which delays the total response of the system and puts up the cost. We also want to allow the user to include any synonyms he needs that have not been presented to him by the system. The system will store these synonyms and at a later stage an indexer can examine them and decide whether or not to link them within the thesaurus. This 'related word' package will necessitate substantial changes to our thesaurus structure and hence to our indexing and retrieval programs. However, we feel the resultant enhancement of our system will be considerable if it does not turn out to be too costly and time-consuming to the user.

(d) Housekeeping

We need the development of a set of basic housekeeping programs so that files can be promptly and efficiently edited on-line and off-line. This will enable us to check for and amend operator errors, to back up the security measures and to allow inspection and modification of files as required. We also need programs to provide an alphabetical printout of the thesaurus indicating which words had been tagged as synonyms and which are related words.

(e) Security

The present security precautions must also be extended and enlarged to include regular selective dumping of portions of files and programs to recreate the system in its latest possible state before breakdown. We will also have to keep a register of input on backing store when our planned off-line mode of indexing becomes operative, to allow for system breakdowns.

(f) Analysis of System

A statistical analysis of our systems will be continued with programs developed to analyse our pilot projects. This analysis has already given us information on methods which will improve the efficiency of our software (see our paper: "Disc Access Algorithms"). In particular, it helped to compress the data within the area where we store records. This study of the basic design of the system will continue throughout, including statistical studies based on new information emanating from the system as it grows. We expect that the results of these technical studies will be continuously fed into the system, improving it and making it more efficient. One of these studies will be a survey of text compression techniques and an attempt to improve on previous methods using a statistical survey of the frequencies of pairs, triads, etc. of characters in the text.

2.2 Query Formulation

As our data base grows larger and new facilities are added the response of the system to the user will change and the query language or command language, as it is variously known,

which was suitable for 1000 documents will not necessarily be suitable for a file of 10,000 documents. Appraisal of the user reaction to the system will be used to decide where it is necessary to modify this system. We are fortunate to have a large group of potential users of our system at Queen's University many of whom are familiar with the use of computers and who work in the area of Atomic and Molecular Physics, including members of our information group working on the Databank supported by OSTI. Even though our initial system may have many teething problems this group of users are likely to give constructive criticisms and so we shall be able to build up a worthwhile assessment of the system.

2.3 Reasons for OSTI support

The above proposals (plus others we will mention in assessing our present position at the end) were acceptable to OSTI. Support was thus given for the continuation of the project with emphasis on the query formulation side in order to allow the information side of the work (as opposed to the computing side) to be strengthened and related to user needs before the system is developed further. It was also felt necessary to evaluate the results of the pilot study in detail in order to demonstrate the applicability of the work outside Belfast and to show that duplication does not occur with other systems.

3. PROBLEMS WITH EXPANSION

3.1 Off-line Indexing

In the early stages of the development of our project we found that the most efficient method of indexing records was in an on-line mode. This was due to the fact that our thesaurus was small and so each record contained many new words. These had to be indexed and added to the thesaurus. However during the past year our data base has increased six-fold, the thesaurus has also grown so that new records now contain very few words which are strange to the system. We have therefore changed our indexing program to work in an off-line mode. Each word in an abstract is indexed automatically if the word has arisen previously, but words not previously in the thesaurus, instead of being displayed on a teletypewriter are stored on a temporary file on backing store.

When a sufficient number have accumulated another housekeeping program retrieves them from their temporary store and is used on-line to record indexing decisions about those new words. The result is that the number of abstracts which can be added to the file each week has considerably increased. The indexer can now make decisions about the significance and stemming of words in his own time and then input them rapidly at the teletype console. This new version of the indexing program has the further advantage that it uses mainly cheaper overnight runs. Indeed the operators can run the program as a background job any time there is 5K of spare core in the machine as the only peripheral it uses after loading is the disc store.

3.2 A Prior Dictionary

At present our dictionaries of words in the data files are built up as the records are updated; to delete all the records currently in the file and replace them with others would require the generation of a new dictionary, the old one being obsolete. In parallel with the maintenance of the present system, we are restructuring the dictionary to allow us to insert words before they are encountered in the updating phase. Thus we can take a set of words and use it for many sets of records for which they are relevant. For example, we could store our present set of Inspec records on magnetic tape, and build a file from the tapes to be issued in 1973 with practically no indexing labour.

3.3 Synonyms

The words in the present dictionary have one tag, the address in their entries; a zero address indicates that the word is non-significant and synonymous words have the same entries address. In the new version we are giving 6 tags (the number is a parameter) to each word. One tag is used to link synonyms in a circular chain and another is used to classify a word as significant but with no entries. The application of the dictionary to a new set of records would only require the automatic resetting of these tags.

The increased number of tags will allow us to include a hierarchical structure if this is considered desirable.

3.4 Housekeeping

3.4.1 Alphabetical printout of thesaurus

During the year the program ALPH was written to give an alphabetical list of all the words in the thesaurus or in any required section of the thesaurus. The location of each word given by bucket and word number is also printed out. For each significant word in the thesaurus ALPH records the number of times that it has occurred and the position of its entry list. Using this program we can detect any errors which may have occurred in the entry lists. An alphabetic listing of the words also shows which words have been linked as synonyms and any incorrectly spelt non-significant words which may be deleted from the thesaurus. ALPH can also be used to obtain a list of the words in the order in which they are stored, i.e. in ascending bucket and word number. Using the program an alphabetic printout of the dictionary as existed in our Inspec disc file in June 1972 was recorded in a Special Report, SR6.

3.4.2 Pre-editing Entry Lists

After studying the QUOBIRD system it became apparent that the speed of the retrieval stage depended to a large extent on the amount of list processing carried out. Therefore if any operation could be carried out at the indexing stage, when time is not as important and the action is only required once, and not at the retrieval stage, when speed of response is important, the system user would benefit. One such case was found after the long-list package (see 1971 Annual Report) was written:- after an entry list has been picked up in the retrieval stage, first it is sorted and

repeats are thrown out, that is, cases where a word had appeared twice in the same sentence in a record. When this operation takes place in core the time involved is not very great, but when the entry list is too large to fit into core quite a number of disc accesses will be needed. In order to avoid this problem a simple house-keeping program was written which checked through existing files, sorted all entry lists and threw out repeats. It then only required a minor modification to the indexing program to check each time a new reference was added to an entry list if the last entry in the list referred to the same record and sentence and, if so, not to update the entry list. In other words repeat headings are never placed in the entry list and so there is no need to eliminate them. It is recognised that the above procedure loses information about the frequency of occurrence of a word, which is important in a weighting system. This difficulty will be overcome in a new version of the system almost complete.

3.5 File Security Problems

In this section we describe some of the file security problems that we currently experience and the efforts we use to overcome them. We find that files get corrupted, either wholly or partially, due to hardware failures, operating system failures, and human shortcomings. It requires a lot of effort to maintain a pure data base and we find that much of our time is spent not in adding anything new to the system but in preserving intact the files already created. We do not feel that for the purposes of this report it is necessary to dwell on the variety of hardware and operating system's software faults that occur and the reasons for their occurrences except to say that the large Fixed Disc Store on which most of our information is stored has been the prime offender.

We feel it is more important here for us to illustrate the kind of file corruption problems we do encounter, whatever their reasons, and to say what we do to remedy them.

To deal with complete file corruption back-up copies of three tapes per file are kept in a separate building from the computer. The frequency of these tape copies depends on the frequency of file updates: a file which is changed every day is copied to tape twice a week, whereas a file which is changed only once a week is copied every fortnight. A more difficult problem arises when a portion of a file, perhaps a single bucket or even a character is corrupted. It can happen that this kind of corruption can remain undetected in the normal course of events for a considerable time, even outlive the grandfather, father, son security precautions we take.

To demonstrate more fully this kind of trouble, we first recall the filing structure and organisation of the system. It is a three level filing structure which is controlled by a directory at the start. One file is used for the augmented catalogue reference records, another for the list of dictionary or thesaurus terms, which may or may not be subject index terms (depending on their judged relevance), and finally one for the set of entries or postings, which stores in a more random fashion, each occurrence of every index term as a catalogue record is being processed.

A particular hazard is when one or two of the entries in the random postings file get destroyed and we have no way of telling this has happened.. Sometimes this type of trouble can remain dormant for a while and then suddenly spread and create havoc. Whilst it is not possible to completely solve the problem we do have a program that has proved successful in detecting the trouble. It works as follows: it takes each index term in the dictionary and checks through its associated entries to test for inconsistencies and looping; it prints an error message if these occur. This program is run at frequent intervals and each time there is a suspicion something has happened to a file. As a further measure of keeping the entries file free of trouble, each time an entry is being included, checks for inconsistencies are made. For example, the last word in every bucket is kept empty and should the program find something in this word an error message is printed out to the effect and the program halted. Again each entry to this data file requires two empty computer words. If the indexing program finds some place on the random file where this condition does not hold then again appropriate action is taken. Another security measure is one capable of dealing with sudden machine

failures, which (unless care is taken) can leave a program in an undetermined state. This can spell trouble, particularly, for our indexing system; built in flags which can be set and then cleared at appropriate points can take care of this trouble. The first thing our indexing program does as soon as it is initiated is to check if such a breakdown has taken place - if so the program then knows to take the necessary steps. This safeguard was described in last years annual report.

A useful aid that we use is to keep one word in every bucket to record the bucket number itself. Sometimes we find that a program can read the wrong bucket and fail to show on the reply word. Thus without realising anything is wrong the program amends the wrong bucket and proceeds to write it back to the "right" place. By having our built-in bucket numbers this error can be controlled. Another useful practice we have included is to keep our retrieval file separate from our updating file. This is done by keeping a file for weekly updates in which we attempt to clear any rubbish before it is merged with the retrieval file when a new weekly file for updates commences. We find that we can detect trouble in the smaller file more easily if it goes straight on to the master file. Finally we have built in the dates of updates to files so that when indexing commences a message is first printed out to say when the file was last processed. This helps particularly when files on disc are recreated by the computer centre whose job it is to manage all application files.

The Computer Centre may recreate a file from its own tape copies after a failure, but not realise that between the time when it took the tape copy and when the failure happened the file was edited. Noting the date of the last update or edit makes sure that errors cannot occur from this source.

4. COMMAND LANGUAGE IMPROVEMENTS

4.1 Introduction

Whilst we are fortunate to have a large group of potential users of our systems at Queen's University, particularly in the field of atomic and molecular physics, we, nevertheless, have to make certain fundamental efforts to induce them to use the system. Our first priority has been to keep the data bases up to date. Also to facilitate users, we have sent out circulars telling them about the service offered and, in addition, we have enclosed simple and pictorial illustrations on how to use the consoles (see Appendix A7). Apart from this we have spent a fair share of our time giving demonstrations and talks on the system. Before enlarging on the improvements we have made to accommodate our user audience, we first give a summary of the information retrieval system using the physics abstracts database as it was at the end of 1972.

4.2 Description of on-line retrieval system for Physics abstracts

ABSTRETR is a system designed to retrieve references on Atomic and Molecular Physics. For 1971 and 1972 the sections taken from Physics Abstracts are:-

13.00 Atomic and Molecular Physics

13.20 Atoms

13.23 Hydrogen and Helium Atoms

13.25 Isotopes

13.30 Molecules

13.31 Inorganic Molecules

13.37 Intermolecular Mechanics

From January 1973 sections 5.2 and 5.4 of Physics Abstracts will be added to the database. At present there are over 6,000 abstracts stored in the system.

In the retrieval program words which describe the subject on which information is required, i.e. search terms or 'keys' are matched against the titles and texts of abstracts. 'Keys' which may be used for retrieving include chemical elements and compounds, experimental processes, mathematical procedures, abbreviations, e.g. RKR, SCF, personal names (where they occur within a title or abstract) and chemical bonds between atoms represented thus: H-C-N.

The linking of words as synonyms is confined to alternative spellings, e.g. sulphur and sulfur, and the symbols for chemical elements with the name of the element itself unless the symbol may also be a common word such as: IN; AS; BE. Pairs of words which are sometimes written with a hyphen joining them and sometimes as one complete word are also linked, but there is no link where the words of the pair are written separately. Thus "auto-ionisation" and "autoionisation" are linked but not "near-threshold" and "near threshold".

To recall the maximum number of documents about a subject the searcher must use 'keys' for all the possible variations in describing the subject. Some information about members of a group may be found by searching under the group name as well as the individual member, e.g. to look under halogens as well as chlorine.

To limit the number of documents recalled the 'keys' chosen should preferably not include frequently occurring words such as electron, atom, molecule but should describe precisely the element, compound and procedure required. The

number of documents recalled by some frequently occurring words can be reduced:

(a) by adding a prefix to the key e.g. L-shell
Pl-electron

and (b) by using the word in a phrase which must occur within one sentence of an abstract to be recalled e.g. auger electron spectroscopy

NOTE If the words of the phrase are used as separate keys the program will find the number of abstracts in which they occur together; the words may then be in different sentences.

4.3 Applied Mathematics, Physics and Computer Science Library

During the past year the books of each of the three departmental libraries consisting of applied mathematics, physics and computer science have been merged. As stated earlier almost all of the 1,000 books in this collection are now in a computerised data base allowing retrieval of information from their titles and chapter headings. However, with the merger, a problem arises because each of the three separate libraries uses its own classification scheme. The cataloguing department of the main library has undertaken to change this to the appropriate Library of Congress class mark to conform with the rest of the main university library. In anticipation of this impending change, the field allotted to the reference or classification number in the record for each book in the applied mathematics and physics books file has been left blank. The correct numbers have not yet been inserted as their allocation is not yet complete. The computer science books, which were processed first, had their own reference numbers. A program, MARK, has been written to put the new class mark in place of the old one.

4.4 Hyphenated Words

The first main change in indexing to be made during the year was the procedure for dealing with hyphenated words. Initially the indexing program treated the words on either side of the hyphen as separate words. This method was not entirely satisfactory if one of these words is non-significant when used on its own. Perhaps the most obvious example of this is the term"on-line" where "on" cannot be indexed as significant. Other problems arose from the inconsistency in the use of hyphens in the text of abstracts, e.g., "ultraviolet" appears in this form as often as in the alternative "ultra-violet". This means that when retrieving records containing "ultra-violet" it would be necessary to use "ultraviolet" and "ultra-violet" as two separate keys. There were so many examples of these problems that it was felt necessary to alter the indexing method. The new version of the program, which stores both words of a hyphen pair, with and without the hyphen, removes the two difficulties described above; the program checks the thesaurus for the previous appearance of each word of the pair and presents them for judgment if this is the first appearance. The words are considered as a pair and also in relation to each other overcoming the problem arising if one of them should be non-significant. As the words are also considered without the hyphen being present the problem of inconsistency is eliminated. This change in the program highlighted the great variety of ways in which hyphens are employed in abstract texts. There are numerous examples of three words joined by two hyphens, e.g. time-of-flight; fuel-to-oxidant. Somewhat less frequent in occurrence are four words joined by hyphens and very occasionally five, e.g. electron-acceptor-electron-donor; valence-shell-electron-pair-repulsion. There are also words followed by a hyphen and without a second word attached, e.g., boron-, carbon-, nitrogen-, fluorine-like, or words with a hyphen preceding them, e.g.,

-inimo or even -/-quark. Complex inorganic compounds usually have names complicated by hyphens as in trans-bis (diphenyl-O-selenolatophenylphosphine). Personal names used to describe procedures are often paired together by a hyphen. This leads to some odd looking "words" in the thesaurus where the names are linked without a hyphen but doesn't affect the retrieval process. This method of dealing with hyphens makes it easier to include bonds between atoms or radicals as retrieval terms, e.g. information about the links H-C-N can be found more precisely than by intersecting H, C and N separately.

4.5 Suppression of superfluous printing during retrieval

After a user has retrieved a set of documents he can ask for either the references, text or sentences to be printed out. There is unfortunately no mechanical way of stopping the program if it is printing out a lot of irrelevant material. A line limit is imposed by the operating system MCS which returns the user to monitor level after a predetermined number of lines and he may then resume the printout at the position at which it was terminated or he may reload the program and start the search again. This is not a very satisfactory solution as most users want to return to the search but omit the printing of the references. It was decided that the best way round this problem was to initiate a counter each time printout of either references, text or sentences was requested. Then every time five references (or five abstracts or sentences) have been printed the user is asked if he wishes to continue the printout, start a new search, exit from the program or return to the stage he had reached in his search before he asked for a printout.

In the near future it is planned to include in the system a special query to the user if he asks for an unusually long printout. This will warn him what he is doing and advise him to have it printed out instead in an overnight run on the line printer (see next section).

4.6 Off-line Printout

As the size of the data-base grew (at present 6500 records), so the number of documents retrieved by any key usually increased also. Whilst some users are prepared to use further keys to narrow their search down to a small number of documents, in other

cases the user may want to retrieve a large number of documents and view them all. As it is very tedious to examine the abstracts of a large number of documents on-line it was decided to provide users with the option of printing their references and abstracts off-line. From the software point of view this entailed providing a special file in which users' names, addresses and a list of the references to be printed were stored. Every night this file is examined and the required references printed out and the file is cleared ready for re-use the following day.

5. ADDITIONAL RESEARCH

5.1 FORTRAN or Assembly Language

The software for the QUØBIRD system was originally written in the ICL 1900 assembly language, PLAN. This can only be used on the ICL 1900 range of computers and is difficult to amend, so as a first step towards re-writing the whole system in the more universally used language FORTRAN it was decided to rewrite part of it and to compare the efficiency of this code with that of the original PLAN programs.

Fortran is a language designed specifically for manipulation of numerical data, and as such does not have any special facilities for storing and handling text or character data. It is not therefore on the face of it particularly well suited to writing programs which involve large amounts of character manipulation. However we overcame this deficiency by storing four characters to one 24-bit word, and treating this as an integer, in CØMPRESS INTEGER mode. We found that character handling was greatly facilitated by the use of two Fortran standard subroutines, CØMP and CØPY. CØMP compares two character strings for equality, and CØPY copies a character string from one location into another.

When we had written the main QUØBIRD software in Fortran, its efficiency was compared with that of the Plan version by setting up a very small data base (6 books on Quantum Mechanics) using each set of programs in turn. We found that the Fortran programs took on average approx. 2.6 times as much CPU time and used approx. 2.4 times as much core storage as the PLAN programs. We felt that this drop in efficiency was probably acceptable as far as the data base generation programs are concerned,

since they are essentially used only once for each batch of documents which is added to the system. It was felt that the loss of efficiency caused by using the Fortran version of this part of the software was offset by the advantages gained, e.g., ease of modification of these programs when they are coded in Fortran.

However as the augmented catalogue retrieval programs are being used constantly, their efficiency in terms of mill time and core area used obviously contributes directly to the overall economic feasibility of the system. We do not therefore think it would be advisable to use the Fortran version of the Document Retrieval program in a commercial working system, because of the resultant substantial drop in economic viability.

Work is also in progress at the moment to write the QUØBIRD data base generation software in another high level language, PASCAL (Reference 1), which has excellent character manipulation facilities. To this end the PASCAL compiler in use at Q.U.B. has been modified to provide a random access file facility and the new PASCAL program is written but not tested yet.

5.2 Efficiency of Hash Indexing

A study is near completion on the efficiency of the construction of the inverted file, with particular regard to how this efficiency varies with the "bucket capacity" used in the thesaurus. A bucket is a block of data which can be written to or read from disc in one operation of 40 keywords. Each keyword is placed in its appropriate bucket by a hash addressing technique based on the division of the binary integer which represents the first four characters of the word, by the number of buckets available in the main storage area. If a bucket overflows the keywords which it contains are split up between it and two overflow buckets by hashing them again, using the number three as a divisor.

since they are essentially used only once for each batch of documents which is added to the system. It was felt that the loss of efficiency caused by using the Fortran version of this part of the software was offset by the advantages gained, e.g., ease of modification of these programs when they are coded in Fortran.

However as the augmented catalogue retrieval programs are being used constantly, their efficiency in terms of mill time and core area used obviously contributes directly to the overall economic feasibility of the system. We do not therefore think it would be advisable to use the Fortran version of the Document Retrieval program in a commercial working system, because of the resultant substantial drop in economic viability.

Work is also in progress at the moment to write the QUØBIRD data base generation software in another high level language, PASCAL (Reference 1), which has excellent character manipulation facilities. To this end the PASCAL compiler in use at Q.U.B. has been modified to provide a random access file facility and the new PASCAL program is written but not tested yet.

5.2 Efficiency of Hash Indexing

A study is near completion on the efficiency of the construction of the inverted file, with particular regard to how this efficiency varies with the "bucket capacity" used in the thesaurus. A bucket is a block of data which can be written to or read from disc in one operation of 40 keywords. Each keyword is placed in its appropriate bucket by a hash addressing technique based on the division of the binary integer which represents the first four characters of the word, by the number of buckets available in the main storage area. If a bucket overflows the keywords which it contains are split up between it and two overflow buckets by hashing them again, using the number three as a divisor.

During the retrieval process, each time the thesaurus file is interrogated means an extra "disc access". And each disc access costs money. It can be seen therefore that the number of overflows in the thesaurus must be kept to a minimum in order to keep the running costs of the system as low as possible.

The mean number of overflows per record in the inverted file was found to depend on three factors: the bucket capacity, the packing density, i.e, the number of records stored as a percentage of the total capacity for records, and the overflow technique. There are a number of methods for dealing with the overflow problem, one of which has already been described. Others include serial overflow (Reference 2) minimum overflow (Reference 3) quadratic overflow (Reference 4, 5, 6) and random overflow (Reference 7). It is not necessary to describe these techniques in detail here, suffice it is to say that they all work on the principle of directing overflow records into a bucket in the same storage area which is not yet full.

For each of these overflow techniques we simulated one hundred inverted file systems, with bucket size running from 1 to 100. When each of a number of predetermined packing densities was reached the number of records which had overflowed up to that point was recorded, and the quantity $\bar{a}$ (mean number of disc accesses required to locate a record in the inverted file) was computed. For these experiments a random number generator was used to simulate the random filling up of the thesaurus buckets with keywords by the hash addressing system. Each experiment was repeated one hundred times, and the mean was recorded as a reasonable estimate of $\bar{a}$ in each case. The probability of error was also determined. The results of this research will be published shortly.

5.3. Machine-Independence of Data Bases and the Programs which Manipulate them

There are two aspects to the machine-independence, or portability of any software system: the portability of the programs which constitute the system, and the portability of the data upon which these programs operate. In the case of an information retrieval system, processing stored data, clearly both aspects must be taken into consideration. During the past few years these aspects of portability have been investigated in a project carried out in the Department of Computer Science at Queen's University, Belfast.

Consider programs which may manipulate data bases, held on auxiliary storage media, in both a sequential and a random access manner. Portability requires that a machine-independent interface be constructed between such programs and the basic facilities for driving auxiliary storage media on all computer systems on which they are to be used. An interface of this nature cannot be completely implemented in a high-level programming language: a small, well-defined set of machine-language subprograms is required for each distinct computer system. In the course of the project, interfaces have been constructed for the ICL 1900 series and IBM System 360 using a judicious mixture of ANSI Fortran and the appropriate machine language in each case. These interfaces have been tested using a locally-developed data processing program, written in ANSI Fortran, the data base being set up from scratch on each computer system.

The complementary problem of data base portability is currently being investigated and takes as its starting point

the existence of a data base manipulated by ANSI Fortran programs, and set up using the interface briefly described in the preceding paragraph. Some additional software is required to convert such a data base into a suitable form on magnetic tape for transfer to another computer system, and to perform the converse operation. Preliminary work involving small amounts of data is being carried out at present, and it is hoped subsequently to set up and transfer a substantial data base meeting the requirements outlined above.

Eventually, when the cost of data transmission over long distances is reduced to an acceptable level, portability of programs and data bases may cease to be of interest to the designers and implementators of information retrieval systems. This will, however, never be true of all software systems, and for the immediate future will also continue to be an important consideration for those involved with information retrieval systems.

5.4 Data Compression

Due to the large number of abstracts being constantly added to the data base, the need for more random storage access is rapidly expanding. It would therefore be advantageous to have some means of data compression to reduce storage costs. This problem of compression has been studied by two research students.

The basic principle involves the substitution in the data of single characters for regularly occurring groups of letters or symbols, e.g. "IONISATION OF THE ATOMS" can be reduced from 23 symbols to 10 as (ION)(IS)(AT)(ION)(OF)(THE

(AT) (O) (M) (S). These groups of characters, composite characters, vary in length from 2 to 16 characters in our system. A subset of the data held on disc, about 200,000 characters,was scanned and 15 lists of around 200 members each were compiled containing the most commonly occurring composite characters of each length. The data was then compressed using various combinations of these composite characters and an estimate of the compression obtained. The storage saving was usually between 27% and 35%. This was disappointingly small and we doubt if the computing time needed to pack and unpack the data from its compressed form is worth the effort.

6. ASSESSMENT OF PRESENT POSITION

In assessing the present position of the reference retrieval system we look separately at the two main supported sections of our work, viz. continuation of the present project and query formulation and below we list some of the enhancements we hope to include in the coming year.

(a) Alternative Retrieval Mode. As mentioned previously, we designed our retrieval system to be self-explanatory as far as possible. While we have found this approach to be very effective with the casual user, we have found that the more sophisticated and frequent users of the system would prefer greater flexibility than is available at present in building up their search queries. They have not, in fact, been hampered in this by the search logic within the retrieval program, but by the natural language in which the search query is formulated. We feel therefore that this type of user should be provided with a different path through the search program to enable him to build up more complicated Boolean expressions easily. For example, we will allow the use of logical and and or in statements, and we will allow a command "STORE N" in which the list of entries at any point in the search can be stored and retrieved later by the command "RETRIEVE N". In this, "N" could take any number between 1 and 8.

(b) Analysis of Systems. A statistical analysis of our systems will be continued with programs developed to analyse our pilot projects. This analysis has already given us information on methods which would improve the efficiency of our software. In particular, it helped us compress the

data within the area where we store records. This study of the basic design of the systems will continue throughout, including statistical studies based on new information emanating from the system as it grows. We expect that the results of these technical studies will be continuously fed into the system, improving it and making it more efficient. One of these studies has already been done in a survey of text compression techniques and an attempt to improve on previous methods using a statistical survey of the frequencies of pairs, triads, etc. of characters in the text (see section 5.4).

(c) Indexing and Stemming. We intend to investigate thoroughly many stemming problems which we have encountered when indexing our records and the feasibility of solutions which have been suggested. These include such items as the indexer needing some form of authority on which to base decisions, for a minimum stem-length being set according to the number of characters in a word, e.g. a word of N characters would not have a stem of less than N-4 characters. We intend, for each subject we are indexing, to provide a list of 'danger' words. These would be words that would always need to be presented in the context of each particular abstract. Such a word which has arisen in our work is AL, which arises as the chemical symbol for aluminium or in the expression 'et al'.

(d) User Manual. Whilst we have designed our system to be as self-explanatory as possible we feel that a sophisticated system will contain facilities which are not apparent to the casual user and for this reason we feel that the

production of some form of manual is an urgent necessity. In addition, this manual would also help the user to understand problems which arise from his interaction with the operating system, e.g. the user could consult it if he has trouble logging in or, for example, if he does not know how to reactivate the teletypewriter if he is timed out, etc.

(e) Language. Whilst we designed the command language of our retrieval program with an outside user in mind, this was not true of either our indexing or housekeeping programs and the command language of these programs needs to be modified to help outside users of this software. It is not self-explanatory: there is no facility by which the indexer can ask for more information and he cannot, for example, see the sentence from which a particular term arose in the abstract or title. The housekeeping programs used for correcting data already on the disc were written for our systems programmers and could not easily be used by an indexer (unless he could write octal instructions!).

Regarding the query formulation side we envisage three main ways in which we will assess user reaction to the system:-

(1) By compiling a questionnaire which our initial "test" users will be asked to complete after each period of on-line searching. We will also approach users personally to discuss their reactions.

(2) To study the conversations between users and the system we will (with the user's permission) duplicate copies of conversations on a special teletypewriter or onto disc. This will require some changes in the operating system to

References:

1. N. Wirth, The Programming Language PASCAL
Acta Informatica 1, 35-63 (1971)

2. W.W. Peterson, Addressing for Random-Access Storage
IBM J. Res. Dev. 1, 130-146 (1957)

3. L.D. Higgins & F.J. Smith, Disc Access Algorithms
Comp. J. 14, 249-253 (1971)

4. W.D. Maurer, An Improved Hash Code for Scatter Storage
Comm. ACM 11, 35-38 (1968)

5. C.E. Radke, The Use of Quadratic Residue Research
Comm. ACM 13, 103-105 (1970)

6. J.R. Bell, The Quadratic Quotient Method: A Hash Code
Eliminating Secondary Clustering
Comm. ACM 13, 107-109 (1970)

7. R. Morris, Scatter Storage Techniques
Comm. ACM 11, 38-43 (1968)

PART B

Project QUODAMP

Page

Publications and Reports

1. INTRODUCTION 1

1.1 Summary of the Problem 1
1.2 Position at 1/1/72 2
1.3 Position at 1/1/73 3

2. PROPOSALS ACCEPTED BY OSTI 4

2.1 Data Collection and Authentication 4
2.2 Critical Analysis of Data 4
2.3 Management of Data 4
2.4 An Intelligent Data System 5
2.5 Query Formulation 5
2.6 General On-line Data Systems 5

3. DEVELOPMENT OF SYSTEM IN 1972 6

3.1 Data Extraction 6
3.1.1 Literature Search 6
3.1.2 Extraction of Data 7
3.1.3 Present Position 8
3.1.4 Specific Problems which have arisen 8
3.1.5 Summary of Present Position 10

3.2 Potential Representation 11
3.3 Retrieval and Manipulation 19
3.3.1 Introduction 19
3.3.2 Retrieval from the User's Viewpoint 19
3.3.3 Curve-fit of Individual Potentials 21
3.3.4 Retrieval from the Systems Viewpoint 23
3.3.5 Program CHAT 24
3.3.6 Program FITS 26
3.3.7 Program CURV 28
3.3.8 Program DEFL 30
3.3.9 Further Manipulation 31

Page

3.4 A Typical User-System Conversation 33
3.5 Miscellaneous Developments 47
3.5.1 Editing Program 47
3.5.2 Checking Program DI23 47

4. ASSESSMENT OF PRESENT POSITION 52

PUBLICATIONS AND REPORTS

1. "Information Storage and Retrieval on Atomic and Molecular Physics" 1st Annual Report to OSTI, 1969.

2. "Automatic Compression of Numerical Information in a Databank" by F.J. Smith and D.C.S. Allison, Datafair '69. Abstracts, 1969.

3. "Interactive System for Storing and Retrieving Interatomic Potentials" by F.J. Smith, paper read at Gordon Conference, July, 1969.

4. "The Semiclassical Inversion of Rainbow Scattering Data" by J.F. Boyle. Special Report, SR1, December 1971.

5. "An Intelligent On-line Data System" Special Report, SR2, April 1971.

6. "TheEfficient Calculation of the Transport Properties of a Dilute Gas to a Prescribed Accuracy" by H. O'Hara and F.J. Smith, Journal of Computational Physics, Vol. 5, 328-344, 1970.

7. "Transport Collision Integrals for a Dilute Gas" by H. O'Hara and F.J. Smith, Computer Physics Communications, Vol. 2, 47-54, 1971.

8. "On-line Data Bank in Atomic and Molecular Physics" by F.J. Smith. Physics of Electronic and Atomic Collisions, VII ICPEAC, 1971, pp.492-96, North-Holland, 1972.

9. "A Data System Centred on Intermolecular Potentials" by J.F. Boyle, paper read at Gordon Conference, U.S.A., August 1971.

10. "Queen's University On-line Data Bank on Atomic and Molecular Physics" Annual Activity Report for the period 1st November, 1969 - 31st October, 1970.

11. "Queen's University On-line Data Bank on Atomic and Molecular Physics" Annual Activity Report for the period ending 31st December, 1971.

1. INTRODUCTION

1.1. Summary of the problem

An ever increasing problem which is facing the scientist of today is the rate at which new technical data is accumulating. This growth makes it gradually more difficult for the researcher firstly to locate any information he requires and secondly to find it in the form he needs. In particular it increases the risk of duplication of research with all the wasted hours that this entails. Four years ago this group began a study of this question by building a databank on atomic and molecular physics. It was our objective to use this as our example to study the problem of on-line data retrieval as a whole.

Because of the size of the field of atomic and molecular physics, it was necessary to restrict our attention to one particular aspect of it, namely interatomic potentials. The importance of these is their close relationship to many of the physical and chemical properties of matter. Our first aim was to build a system which could store these potentials and to enable them to be retrieved "on-line".

As the system was being built, it soon became clear that it could be developed one stage further from being merely one for retrieval of information to being one which allowed manipulation of stored data. A simple example of this manipulation is to provide the facility for the potentials to be retrieved in whatever units required.

The idea of manipulation can however be developed much further. When a scientist requires a potential his interest is not so much in the potential itself but rather in using it to calculate parameters which define one of the physical or chemical properties of matter. An obvious step then is to build into the system programs which will enable him to carry out such calculations. To do so there are two requirements. Firstly, the programs required for these

calculations must be built and added to the system. Secondly, to enable calculations to be made with a particular potential, an automatic procedure will be needed to construct from the various approximations of the potential which have been stored an estimated potential which is representative of the most reliable results available. The first part of this presents no major difficulty but the second does pose serious problems which are discussed in Chapter 3, section 2.

1.2 Position at 1/1/72

By the end of 1971 several thousand papers had been examined and out of these some 750 which were considered relevant had been located and photocopies of them obtained. Many of these were found on close study to contain data which was not worth storing, for a variety of reasons, e.g. no range given on data replaced by more accurate later data. In general, less than one paper in three was found to contain data worth storing.

Approximately 500 assorted worthwhile potentials were contained in the data bank together with an estimate of the short range potential for the ground-state of every possible pair of atoms which does not include a hydrogen atom; that is, a total of 5,500 potentials in all.

An updating system was in existence which enabled new potentials to be added to the system and to take their correct position with respect to those already in the bank.

A smooth representation of the potential over the whole range is necessary before manipulative facilities of any complexity can be offered. By the end of 1971 a method existed within the system which fitted an analytic function with coefficients chosen to minimise the least square error to a function of the logarithm of the potential. However the results of this, whilst fairly satisfactory, showed that there was still some way to go to solve this problem with precision.

In the retrieval system a user could retrieve any of the stored potential estimates in units of his own choice, with or without the associated references, and with the estimated accuracy and range of validity along with any other comments included about the potential.

Two other options were also available - the first calculated and displayed the most accurate values of the stored potential in a range specified by the user, while the second calculated a representation of the potential over the whole range. In both cases the user could obtain a pseudo-plot of the values on his remote terminal. Alternatively, if he wanted a graph of the potential curve, he could store the data for later output to a graph-plotter, which cannot be accessed directly from the on-line terminal. Whilst the system at this stage was fairly flexible in what it offered, even to the extent of including novel features like on-line pseudoplots, it nevertheless had a big drawback in that it did not permit the user to manipulate with whatever potential he chose himself.

1.3 Position at 1/1/73

By the end of 1972 almost all of the possible relevant papers have been located. The reading of these papers and the extraction of interatomic potentials from them is now virtually complete. A test of the thoroughness of our search revealed that as many as 94% of relevant papers have been traced so far, this being before the data base is even completed. The data extraction is dealt with in greater detail in section 1 of Chapter 3.

The present report describes the retrieval side of the system both from a user's viewpoint and from a system's viewpoint. We recommend readers who are only interested in the facilities which the system offers to read the first section as well as the conversation display illustrating what the system does. However, those interested in a detailed description of the software involved should also read the section on retrieval from a system's viewpoint.

Whilst we feel we have added a greater degree of user flexibility to the system over the past years and have incorporated further manipulation facilities into it we realise that the user conversational language still requires some tidying up: we now believe the stage has been reached where this can be looked at more closely.

2. PROPOSALS ACCEPTED BY OSTI

Prior to the ending of a previous grant, proposals were submitted by us for a continuation of the work involved for a further two years from the beginning of 1972. A summary is given in this section of the parts of this proposal which OSTI agreed to support.

2.1 Data Collection and Authentication

By the end of the proposed two year period of the grant the data base of interatomic potentials is to be completed together with a critical evaluation of the data. A system to ensure the data is kept up to date is to be set up. The data base will be broadened with the inclusion of subsidiary data files in subject areas closely related to interatomic potentials. These include

(1) Oscillator strengths
(2) Energy levels
(3) Transport properties
(4) Polarizabilities

A start is to be made on building a new data system based on wave functions.

2.2 Critical Analysis of Data

The evaluation of the data will be carried out in two stages. Initially this will be carried out by us based on our own experience. However outside experts employed on a consultancy basis will be used to consider this problem as a whole when the data base is complete.

2.3 Management of Data

Programs to enable alterations to be made to the stored data are to be written. A number of housekeeping programs will be needed to implement security procedures and also to

keep accounts of the use of the system.

2.4 An Intelligent Data System

Programs will be written to enable the user not just to retrieve the data but to manipulate it, that is, to compute other data with it, to change its form, to generate related data, etc. Programs suitable for inclusion will have to be tested for the accuracy of their results and modified to a form compatible with the system.

2.5 Query Formulation

It is essential to adapt the system to the needs of the user who may either be experienced in using it or just a novice. To this end, it is proposed to develop two systems, one for each of these users.

Some means of obtaining the user reaction to the system will be developed in order to judge the success of it.

2.6 General On-Line Data Systems

An investigation as to the viability of applying the general techniques adopted in the present project to an on-line data system in other fields such as the medical and engineering sciences was proposed. However OSTI felt they could not support this at this stage but may do so at some later date.

3. DEVELOPMENT OF SYSTEM IN 1972

3.1 Data Extraction

3.1.1 Literature Search

The potentials which form the data base are found in papers published in the various scientific journals. The five different methods of searching the literature were considered in detail in the 1971 Annual Activity Report and are listed here without comment:-

(a) "Physics Abstracts": published by Institute of Electrical Engineers
(b) "Bibliography of Atomic and Molecular Processes": published by the Atomic and Molecular Processes Information Centre at Oak Ridge National Laboratory
(c) Review Papers and Books
(d) Personal Communications
(e) Literature references

By (e) we mean references within one paper to other papers which may also contain relevant potential data. This continues to be a very important source of references, particularly to papers dating back to before 1965.

By the methods outlined in the previous paragraph some 1330 papers have been located and photocopies of these made. This includes all the relevant papers up to the end of 1972 as reported by "Physics Abstracts".

An important question that obviously arises is how thorough a search has been made. Whilst one can be confident of an almost 100% success rate for publications in recent years (say since about 1966) one can be less certain about earlier papers. This is supported by our experience as we read through the papers which indicates that it is the older publications which are harder to trace. Gradually, however, these gaps should be filled in as the search proceeds.

A more specific indication of the thoroughness of the search was obtained from the references quoted in "A Bibliography of ab initio Molecular Wave Functions" by W. G. Richards, T. E. H. Walker and R. K. Hinkley. This book, published in 1971, lists the references for the best available ab initio calculations for the potential energies for all atomic pairs. Of 154 references made to 107 different papers it was found that only ten had not been located by us.

In an attempt to find ways of improving the efficacy of the literature search, an examination of the ten missing papers was valuable. Four of the papers were dated before 1965, the date from which we began our detailed literature search. As indicated earlier our hope is that any relevant early papers are gradually found, as indeed these four have been, through references to them in more recent publications. This shows that only 6 papers out of 107 were not found which indicates a 94% recall at a time when the data bank was not complete.

3.1.2 Extraction of Data

The actual data which is extracted from a particular paper has been discussed in some detail in previous Annual Reports and a list is given here without comment:-

(1) Paper title and reference
(2) Diatomic system and state name
(3) Type of potential
(4) Method of calculation
(5) Parameters, formula or values which actually define the potential
(6) Units
(7) Error and source of error
(8) Range; numerical and indication as to

whether it is short, intermediate or long range

(9) Any other relevant information.

A number of other items of data peculiar to particular types of potential has been added to this list. Firstly, for potentials calculated using the Rydberg-Klein-Rees (RKR) method it is essential to extract D_e, the well depth of the potential concerned and T_e, the height of the potential minimum of the state being stored above the potential minimum of the ground state. Secondly, numerical results obtained from ab initio calculations are mostly given as total energies rather than potential energies. This requires that the separated atom energies of the atom pair concerned be stored for numerical potentials. In cases where the author already has changed from total energies to potential energy a value of zero is stored instead of a separated atom energy.

3.1.3 Present Position

Of the 1330 papers which have been located approximately 1250 have now been read. From these data has been extracted from 340 papers, which is about 25%. A total of 200 papers have been processed only partially since they present problems which have as yet not been resolved (see next section). The number of papers which have not yet been read is small and it is expected that these will be processed in weeks rather than months so that the data base will be up to date by February 1973.

The data from some 200 papers is at present being tested prior to being added to the potentials already in the data bank.

3.1.4 Specific problems which have arisen

(i) As yet we have no means of storing potentials which have been expressed by the author in the form of an analytical expression. The forms chosen rarely fall

into a general pattern and the only solution may be to store each one (of about forty to fifty) separately as if it were a different type of potential.

(ii) A large number of potentials are presented in a graphical form and as yet we have tried no method of reading these accurately. Indeed many of the graphs given are so small as to make this task extremely difficult. Nevertheless, with some sixty to seventy papers with their results presented in this form, this forms a large quantity of at present unstored data.

(iii) It has been found that in forming potentials of a parametric nature (e.g. of Lennard-Jones, Kihara, Morse and Buckingham types) the authors rarely give the range over which they regard their potential to be valid. Since our programs require this to be specifically stated this has meant the choice of some range based on experience. We have decided on the range

$$0.9r_m < R < 3.0r_m$$

where r_m is the internuclear separation at the equilibrium point. Nevertheless, this choice is still arbitrary and we feel the whole question to be one which would be well worth a much fuller investigation.

(iv) It is sometimes not possible to obtain the values of D_e, the potential well depth, which is required for storing RKR potentials. This also applies to the value of the separated atom energies which is required for many numerical ab initio calculations. These missing values will have to be added into the databank as and when they become available.

(v) The most difficult problem which is met in building up the data base is still the evaluation of the data; that is, estimating the accuracy of the values given for the potential. The need for some sort of evaluation is readily recognised if the user of the system is to be able to discriminate between a number of estimates of the same potential. However, since in most cases the authors of papers do not give any indication of the accuracy of their results, it is left to us to provide an initial evaluation. This is based on a number of

factors; the method of calculation used and a comparison with results from other sources being the most important. Often this approach is adequate but at other times it merely constitutes an educated guess. However, with the completion of the data base now near, we will soon be able to improve on our present methods of evaluation, firstly by carrying out more general and widespread comparisons and secondly by employing experts to look at the particular potentials and so carry out a more realistic assessment of them.

3.1.5 Summary of present position

Summing up we can say that the creation of a data base of interatomic potentials is now almost complete. Once this is done it remains only to ensure that the system is kept up to date with new publications. Attention can then be turned to those other quantities such as oscillator strengths, polarizabilities and energy levels which are to be stored so as to broaden the data bank.

3.2 Potential representation

before manipulative operations of any complexity can be carried out on the stored potentials a smooth representation of them over the whole range is necessary. The aim is to set up a system whereby this curve can be obtained from the available data describing the intermolecular potential for any state of any diatomic system.

The available data may be any combination of the following four:-

(1) The asymptotic form as the internuclear separation R tends to zero which is given by

$$V(R) = \frac{Z_1 Z_2 e^2}{R} + a_o + a_2 R^2 + a_3 R^3$$

where a_o is the united atom energy and Z_1, Z_2 are the atomic numbers concerned. (Buckingham [1958]).

(2) A short-range part, generally fitted by a Born-Mayer potention

$$V(R) = Ae^{-bR}$$

the accuracy of which is poor, particular as R increases.

(3) A set of points distributed around the potential minimum, some of high accuracy (usually those calculated using an RKR procedure) and some of fairly poor accuracy.

(4) The asymptotic form as $R \rightarrow \infty$ usually expressed either in terms of a Van der Waals Coefficient or, more generally, as a long range expansion in inverse powers of R.

As a starting point, it was assumed that the results using an RKR procedure would be available and these would be used to represent the potential in the intermediate range region.

The first attempts were directed at fitting an analytic form(with a maximum of three parameters) to the RKR points. This would then be joined, by means of other simple parametric functions, on the one hand, to the Born-Mayer potential, and on the other hand, to the long-range asymptotic form. For the first part of this scheme, a program was written to fit in turn each of six parametric potentials to the RKR points, choosing the best or ending the search when the average percentage error in the fit was less than 1%. Almost invariably the best fit was given by the Levine potential:

$$V = E\,X(X-2) \quad , \qquad X = \frac{R_m}{R}\; e^{-a\left(R^p - R_m^p\right)}$$

In connection with the parameters, it may be noted that $V' = 0$ when $X = 1$, i.e. when $R = R_m$, and then $V(R_m) = -E$.

Of the different forms tried for the joins, the most successful were:

$$V = e^{-bR}\left(a_0 + a_1 R + a_2 R^2 + \cdots\cdots + a_5 R^5\right)$$

for the lower join, where b is the Born-Mayer value, and

$$V = R^{-6}\left(b_0 + b_1 R + b_2 R^2 + \cdots\cdots + b_5 R^5\right)$$

for the upper join.
Each of these contains six variable parameters, which are utilised to impose continuity of the potential and its first two derivatives at both ends of the join.

The success of these attempts is judged by the degree of "smoothness" achieved, a necessary (though hardly sufficient) condition being the absence of turning points in the join regions. If the join is not sufficiently smooth we might make the trial form more flexible by adding more variable parameters. However, in general more parameters mean a greater likelihood of oscillations, and it is difficult to see how the parameters could be chosen expressly to eliminate these. The most probable cause of lack of smoothness is an essential incompatibility between the different segments of data as given, and alterations to one or more of the segments may be necessary.

The position is rather different with respect to the two joins. If the RKR data is inconsistent with the longrange asymptotic form, it means that we are applying the latter at distances which are too small. We can move out the upper end of the join, which was quite arbitrary in any case. At the lower join, the Born-Mayer and RKR potentials may be clearly inconsistent, indeed, in extreme cases, they might cross. The Born-Mayer potential must be altered in such a case, but we are not free to move in the lower end of the join without limit. It would be better to make some slight alteration to the Born-Mayer potential as a whole, but because of the difficulty of doing this in a meaningful way, a new approach altogether in which the Born-Mayer potential plays almost a secondary role seems advisable. The opportunity is taken at the same time to extend the fit down to $R = 0$, with the aid of what we know about the asymptotic form there.

For the moment, let us retain the parametric potential fitted to the RKR points. We seek a parametric form to be fitted to the whole region interior to this, and another in the whole region exterior to it, thus reducing the number of "pieces" in the potential to three. In the interior region we might try

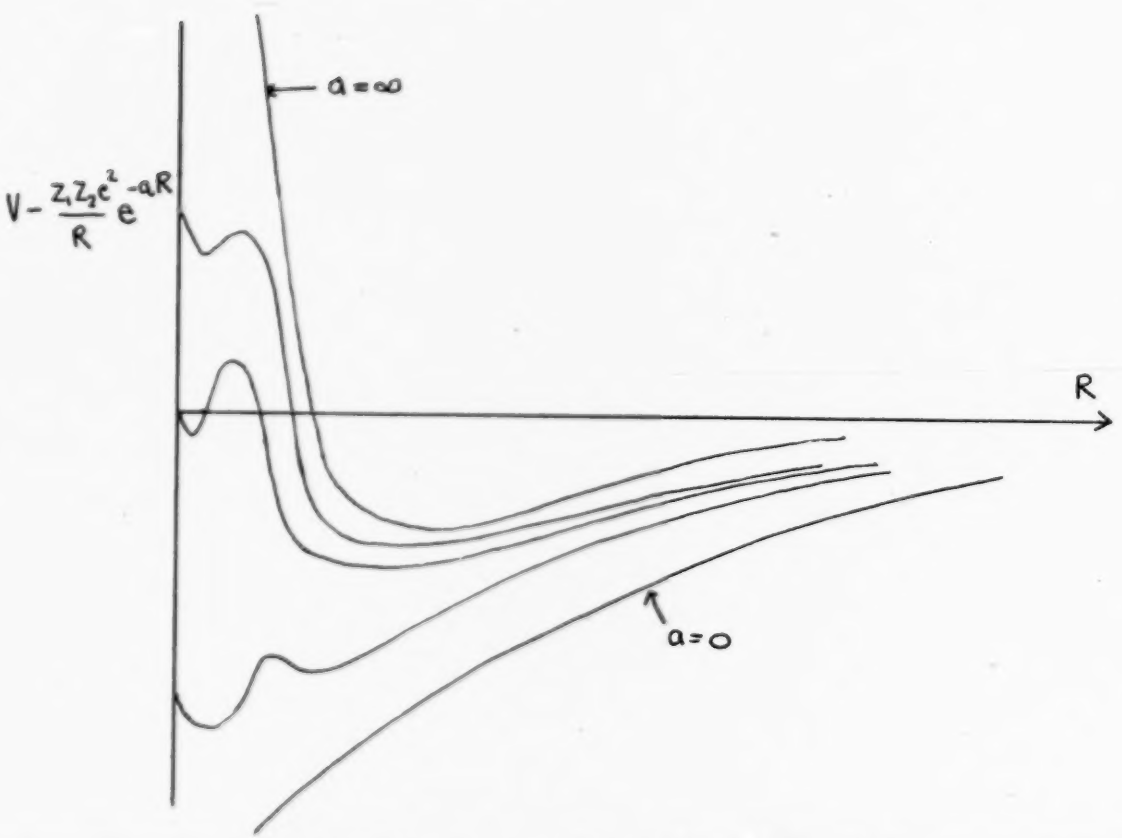
$$V = \frac{Z_1 Z_2 e^2}{R} e^{-aR} + e^{-bR} \left(A_0 + A_1 R + A_2 R^2 + \cdots \right)$$

The significant fact here is that

$$V - \frac{Z_1 Z_2 e^2}{R} e^{-aR}$$

has a finite limit as $R \rightarrow 0$. The figure below plots this function for several values of "a", though the behaviour for small "R" is speculative, because in this region we have only the Born-Mayer data as a guide. We see that if "a" is too small, the subtracted term tends to swamp the original potential over too much of the range, whereas if "a" is too large, the usefulness of having a finite intercept is lost. Comparison with the Buckingham expansion shows that if "a" is properly related to the united-atom energy, the intercept on the V axis is zero, and the A_o term is then unnecessary in fitting the residual potential. However, this fact is not of much practical use, since:

(a) the particular residual potential which goes through the origin appears to cross the axis at least twice first;



(b) united-atom energies are not available for many cases of interest;

(c) even where the procedure is practicable (e.g. Li_2), theoretical expectations were not realised.

We might do better, therefore, to choose any suitable value for "a", and retain A_o amongst the parameters.

The parameters in the residual potential would be chosen with a view to:

(1) Continuity of V and its derivatives at the lower limit of the RKR region;

(2) Approximating the residual part of the Born-Mayer potential over the appropriate range;

(3) The general requirement of smoothness.

Applying the same ideas to the exterior region, we might try

$$V = \frac{C_6}{R^6}\left(1 + \frac{B_1}{R} + \frac{B_2}{R^2} + \cdots\cdots\right)$$

choosing the parameters for continuity with the RKR potential, and for smoothness.

There are great difficulties in enforcing all of the above criteria, particularly that of smoothness. The criterion of continuity may be dealt with by subtracting

$$\frac{Z_1 Z_2 e^2}{R} \quad e^{-aR}$$

from the various data over the whole range, and then trying to fit something over this <u>whole</u> range. The residual potential is everywhere finite, and the interval can be changed to $(-1., +1)$ by the transformation of independent variable

$$t = \frac{R - \alpha}{R + \alpha} \qquad (\alpha = R_m, \text{ probably})$$

We may then fit the transformed data with Chebyshev polynomials; but the great difficulty will be in reproducing the correct long-range asymptotic form when this expansion is transformed back to the infinite interval.

A possible way around this is to make a further addition to the potential:

$$V - \frac{Z_1 Z_2 e^2}{R} e^{-aR} + \frac{C_6}{R^6} e^{-\frac{b}{R}} \qquad \left(\text{or } e^{-\frac{b}{R^2}} \text{ if next term is } O(R^{-9})\right)$$

"b" is not to be so large that the residual potential is still comparable to V for unsuitably large values of R, nor so small that V is seriously distorted in the region of the minimum. We may now cut off the derived potential at a value of R for which it is very much less (in absolute value) than V, apply Chebyshev polynomials in this finite interval, and add back the two terms after the fit has been made.

We have so far outlined three methods of approach which deserve further investigation:

(1) a fit in three pieces, i.e. interior region, RKR region, exterior region.

(2) a fit in one piece, consisting of

$$\frac{Z_1 Z_2 e^2}{R} e^{-aR}$$

plus Chebyshev polynomials over an infinite interval.

(3) a fit in one piece, consisting of

$$\frac{Z_1 Z_2 e^2}{R} e^{-aR} - \frac{C_6}{R^6} e^{-\frac{b}{R}}$$

plus Chebyshev polynomials over a finite interval.

A fourth possibility, going in the opposite direction, is to adapt the Levine potential, which is so successful in representing the RKR points, so that it has the correct asymptotic form at both ends of the range. As it stands, it goes to R^{-2} as $R \rightarrow 0$, and falls off exponentially for large R. If we replace the previous definition of X by

$$X = \frac{R_m + \alpha\sqrt{R_m}}{R + \alpha\sqrt{R}} \cdot e^{-a(R - R_m)} + \frac{C_6}{2E\,R^6}\, e^{-\frac{\sigma}{R}}\left(1 - \frac{R_m}{R}\right)$$

we have the required asymptotic forms. Moreover we have taken trouble to preserve the interpretation of R_m as the value of R for which $X = 1$, which is important if initial estimates of the parameter R_m are to be accurate. α is chosen to give the correct behaviour for small R, and is seen to depend on "a". The least squares fit to the RKR points is now performed with respect to the parameters E, R_m and the curvature at R_m. The expression for the curvature depends also on both "a" and α, and so, for a given curvature, we must find "a" and α by solving this simultaneously with the condition imposed by the short-range asymptotic form. On the other hand, the constant σ may vary within quite wide limits without invalidating the asymptotic form. In assigning it, we may consider

(a) conditions similar to those imposed on "b" on the previous page;
(b) the known coefficients of R^{-7} or R^{-8};
(c) making it a parameter in the fit to the RKR points.

To summarize, the overall purpose of these approaches to the potential representation problem has been to find some modified form of the potential curve which permits an accurate and realistic curve fit to be carried out. As yet this goal has not been achieved. To cover the

possibility of this not being done successfully a less subtle form of potential representation based on an interpolation procedure will have to be implemented.

3.3 Retrieval and Manipulation

3.3.1. Introduction

To understand more fully the characteristics of the retrieval system and to appreciate the changes and improvements made over the past year, the reader is referred to the previous annual report. It is not the intention here to report what has been said in it except where it is felt necessary to clarify differences that have taken place during the interim period. The system's philosophy remains unchanged and is ably illustrated in the following extract:
"Generally when a scientist retrieves a particular potential he is not so much interested in the potential for its own sake, but rather as a means to an end. Usually he will use it to calculate parameters defining one of the physical or chemical properties of matter. An obvious step is therefore to build into the system the programs which will enable him to carry out such calculations."

With this objective in mind the system throughout the past year has incorporated into its conversational mode more user control of the data involved and more emphasis has been put on the system's manipulation capabilities.

3.3.2. Retrieval from the user's viewpoint

We now describe the present state of the system. It differs from that of a year ago in that then the user had no control over which potentials were used in the manipulation facilities; now the user can exercise direct control if he wishes. For example, he could manipulate with a single stored potential when before he had to accept the system's curve fit.

At the start of a search preliminary information is given to the user if he is not familiar with the system. The user then specifies a pair of atoms. Unless he is

interested in only the ground state, which he indicates by following the atomic symbols with an "X", the system gives him a numbered list of the states for which potentials are stored as well as the number of potentials for each state. The user identifies by number his choice of state. If there are no states for the user's atom-pair, he is given the option of the Born-Mayer (B-M) short range potential; the B-M is only available for a ground state and is not possible if either atom is hydrogen, in which case he may then try with another atom-pair. After choosing a state the user then specifies the energy and length units in which he wants his potentials and in return he is given two options:

(i) the ability to choose any of the stored potentials

(ii) the generation of a potential over the whole range obtained by fitting a smooth curve to the stored potentials.

For the first option the user is given a numbered list of the potentials - each potential consists of the potential type, the range over which the potential is valid and whether this range is short, intermediate or long.

The user is now invited to select from the potentials displayed. This selection process is done by choosing one potential at a time. Subsequent to picking a potential by its corresponding number, the user is offered any combination of the following three facilities;

(A) A tabular listing of the potential points

(B) the potential points stored as part of those to be used for further manipulation purposes, and currently this is a curve fit for the potential over the whole range

(C) an off-line graph plot of the potential.

If a listing is required the user is asked for the number of points he wants and the range over which he wants them and whether he wants a short or comprehensive

description of the potential.

For (B) and (C) the appropriate points are written to pre-determined areas of the disc. In the latter case the values are later retrieved in the off-line graph plotting program, while in the former they are used in the manipulation program. If the user asks for both (B) and (C) the two operations are carried out independently.

If the user requests any off-line graphs during the process of this search then he must give the system his name and address for later identification of the graphs.

33.3 Curve-fit of individual potentials

For facility (B) above and the earlier option (ii) the respective individual potential fits are transferred to a curve-fitting package which fits a smooth potential through all the points included, reducing the potential to an equation with four parameters which should be valid over the whole range. As was described in Section 2 the present curve fitting program is not entirely satisfactory and during the past year much effort has gone into an attempt to help to produce a smoother representation over the whole range.

When the smooth representation of the potential has been determined, the system offers the user the following:

(a) a tabular listing

(b) further manipulation possibilities.

The tabular listing is similar to that outlined above for the individual stored potentials; the difference now being that the user is not confined to asking for a number of points within a specified range but theoretically he can choose any range from 0 to ∞ . Having specified his range of interest and number of points he can have them listed and he can obtain a variety of off-line graphs.

Firstly, he can have a curve fit of the potential over any part of the range as illustrated in fig. 1 and marked (1).

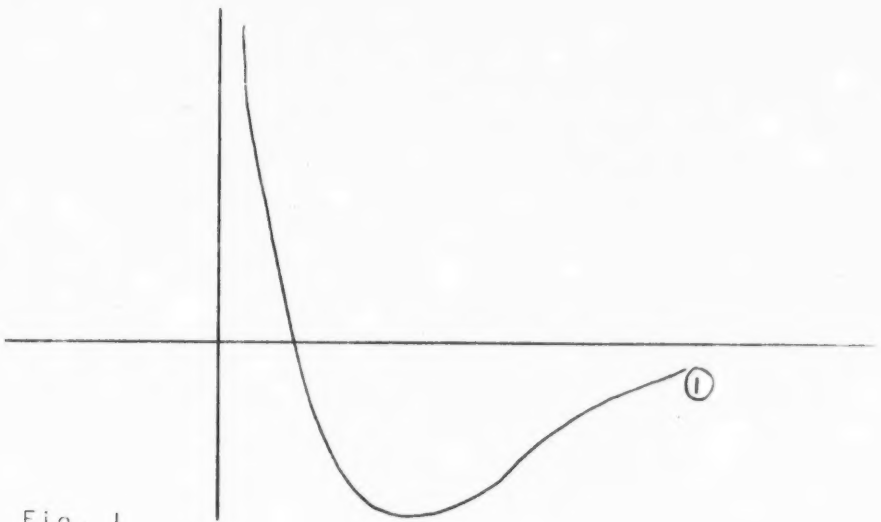


Fig. 1

He can have the same again but this time with all the potentials used in finding (1) as illustrated in fig.2.

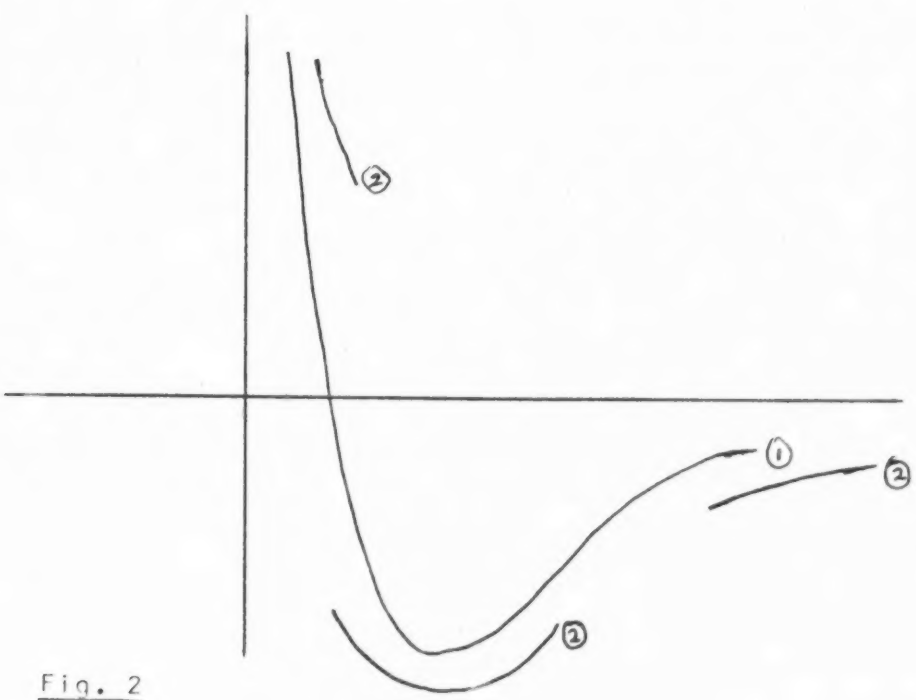


Fig. 2

3.3.4 Retrieval from the systems viewpoint

The retrieval system works in a time-sharing, batch-processing multi-access environment and as such there is a core restriction (at present 18K (24-bit) words) available to individual on-line programs. It is expected that a new 1906S ICL computer (presently being installed) will be operational in the coming summer months. The present machine (a 1907 ICL computer) will then be used mainly for multi-access work and so core restrictions will be greatly improved. Because of this core limitation it is necessary for the whole retrieval package to consist of four individual programs. Only one of these programs can be in core at any one time and so at an appropriate point each (with the aid of a system command given by the user) has to activate the next program. Information between one program and another is conveyed via a communications file which is held on disc as a temporary storing place for the required data. The four programs involved are:

CHAT: the controlling segment which outputs states for a user's atom-pair, as well as transferring to another program, FITS, the user's choice of state and the starting addresses of the associated potentials and units required, via the communications file.

FITS: This program outputs the stored potentials for a chosen state as well as transferring to #CURV the selected potentials for curve fitting the potential over the whole range and storing values for individual graphs.

CURV: This program does the curve fitting of the potentials transferred from program FITS as well as transferring the parameter equation values calculated for the curve over the whole range to program DEFL when further manipulation is required.

DEFL: This program uses the four parameters transferred from # CURV to determine the deflection angle for any energy and impact parameter given by the user.

3.3.5 Program CHAT

The first thing performed in # CHAT is to discover if the user is starting his search or returning from another program. To determine this we call S/R OPEN where we open our communication file INDXABbFO112 and make the appropriate check. If the user is returning to CHAT we must reassign values to his "atom-pair", "statename" and so on. As indicated earlier whenever we go from one program to another, we must write this information (atom-pair, etc.) to disc and then read it down again whenever we have entered the required program. If we are returning from # CURV or # FITS, the name of our data file is stored on the communication file. However, assuming the user is just starting his search, we call S/R INITIAL. In S/R INITIAL we ask the user if he is familiar with the system and, if he is not, give him preliminary information about search strategy and so forth. In INITIAL we may also set switches to give intermediate values in various S/Rs; these are used to help the systems designers to detect errors. INITIAL itself calls S/R PRELIM, which reads from the data file information about potentials and addresses on disc of atom-pairs. We now return to the MASTER.

It is here that the search really begins. The user is asked to type his "Atom-Pair?". The address of this question within the context of # CHAT is stored on disc by calling a plan S/R STORE. This is to allow the user the facility of restarting his search at any time by simply typing "A" for any user response. The user's choice of atom-pair is read in S/R ATOMNOS. ATOMNOS checks if the chemical symbols are valid and determines the corresponding atomic numbers from a preset list. It also calls S/R NAMES which checks for ". + or -" etc., for a user may type HH+, H-H; he may in fact follow his atom-pair with an "X" which tells the system that he only wants to deal with the ground-state, whereupon a pointer is set to indicate this.

Back in the MASTER another pointer is set to indicate if either atom is hydrogen. We next call S/R SEARCH1 which determines the position on disc of the 1st state for these atoms. If there are no states stored for a particular atom-pair we give the user the option of the BORN-MAYER potential; this potential can always be generated by substituting in a set equation the values of certain parameters. Assuming there are states stored the program calls S/R SEARCH2, which lists out all the states in the databank for this atom-pair, with the number of potentials for each state. The user is then asked for his choice of state. After making his choice the program determines the address on disc of the first potential for this state and returns to the MASTER. If he does not want any of these states, he may try for another atom-pair. Next we ask the user for the energy and length units in which he wants his data.

The data on the disc is stored in atomic units and the factors necessary to convert the values to the users units are found in the same preset list as for the atomic numbers. This is done in S/R UNITER. In S/R CLOSEN the communications file is now updated with the atom-pair, statename, units, etc. and all files closed. The name of the data file is also stored in

the communications file. Then control is passed to another program in the system (# FITS).

On returning to # CHAT from some other program in the system the relevant information is read from the communications file and the data file opened. The user may then try another state for the same pair of atoms and have the states listed again if he wants. Failing this he may then try another pair of atoms or cease execution of the program. When given the choice of another pair of atoms, the user is first asked "Other Atoms?". In reply he need not type "Yes" or "No" but may type immediately the symbols for his atom-pair. Similarly in reply to "More States?" he may type the number of the state he wants.

3.3.6 Program FITS

As # CHAT retrieves states for a chosen atom-pair, so # FITS retrieves potentials for a chosen state. The first thing done is to reassign values to the atom-pair, statename and number of potentials for the state chosen. This is done in S/R OPENI, which also calls S/R PRELIM to read information about the potentials. Furthermore, it calls S/R SETGRAPH which sets up the graph and manipulation buckets and determines if there is room available in the graph area. The user is then asked "All or Best". Here he has the choice:-

ALL:- He may have all the potentials listed and choose whichever he wants as described below.

BEST:- Here all the potential values are written to disc for use in the curve fitting program. In this case the user has lost control of individual potentials though in # CURV he may obtain a graph of all the potentials together with the fitted potential.

The potentials are listed in S/R DATAREADI and the user is asked to choose one. If none of the potentials listed

is suitable to the user, he may return to # CHAT to try another state for his atom-pair or indeed a different atom-pair. Provided the user chooses a potential, he is then asked "LIST, STORE, GRAPH?" which means:-

LIST:- does he want the values for this potential listed?

STORE:- does he want these values stored for further manipulation?

GRAPH:- does he want a graph of these values?

The user response to these takes the form of three integers, each taking the value 0 or 1. The latter indicates that the user is interested in the appropriate option. The user can have any combination of these three possibilities. We now call S/R DATAREAD1 which this time will output the users chosen potential. DATAREAD1 will itself call S/R ABRAHAM3 to output the short Born-Mayer potential, S/R DATANUM to output the intermediate RKR potential, or S/R RETC6 to output the long C6 potential. In each of these S/Rs the values will be listed, stored for manipulation, or stored for a graph, depending on the user's responses to the options above. Furthermore DATAREAD1 will call S/R DECODE if the user requires a description of the potential; DECODE will give, for example, author and reference of the paper from which the data was extracted, the amount of information given depending on the user's answer to "SHORT?". DATAREAD1 also calls S/R ERROROUT which gives what is considered to be the relative or absolute error of the data.

Back in the MASTER we ask the user if he wants another potential (Yes, No or actual number of potential). If he types "Yes" or the actual number we go back for another search. If the user does not want another potential, we then check if he wants a graph, etc. Should he want a graph we call S/R STORENAME, which will ask for his name and address (to identify the graph) and write this information to disc. If he wants the values

stored for manipulation, we call S/R CLOSENI to write the communication bucket to disc and then prompt him to type "£RUN,CURV"; otherwise we again call CLOSENI but this time ask him to type "£RUN,CHAT".

Finally a word about LIST, STORE and GRAPH in the context just described. A user can always have the values listed out. He need not limit himself to having a graph of some of the potentials or storing them all for manipulation; he could for example get a graph of 1 potential and store 3 potentials for manipulation. At the moment further manipulation can be accomplished in the calculation of a smooth potential over the whole range and then the determination of deflection angles. More description of these facilities will be given in the following sections.

3.3.7 Program CURV

As in #CHAT and #FITS the communications file is opened and a check is made of the number of users with data already stored for graphs - there is a maximum of 4 put on the number available at any one time. We then call S/R READDOWN to read from disc the values stored for manipulation. Next we call S/R FITS, which curvefits the values using S/R MA02A*; this gives us four parameters which, when substituted in an equation, provides values for the potential (V) for every value over the range (R) from 0 to ∞ (Fig. 1).

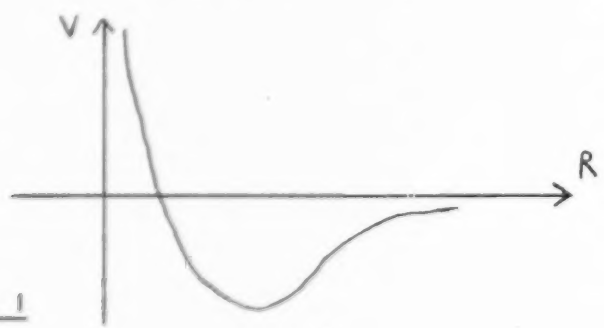


Fig. 1

*S/R MA02A is an Atlas routine which solves a set of linear simultaneous equations.

It is then necessary to know if the user wants values listed in a certain range or if he is interested in deflection angles. In the latter case the user is prompted to pass control to # DEFL. If, however, he wants a list of values, he is asked for the range he requires and the number of points in this range. The values for the potential are then generated at these points and the results listed out. Having studied the values the user may reply in three ways to the question "GRAPH?".

1. If the user decides that the values are unsatisfactory, he should type 0.
2. If he types 1, he will obtain graphs of the stored values and the curvefitted points on the one frame. (See Fig. 2)
3. If he types 2, he will obtain a graph of the curvefitted potential alone. (See Fig. 3)

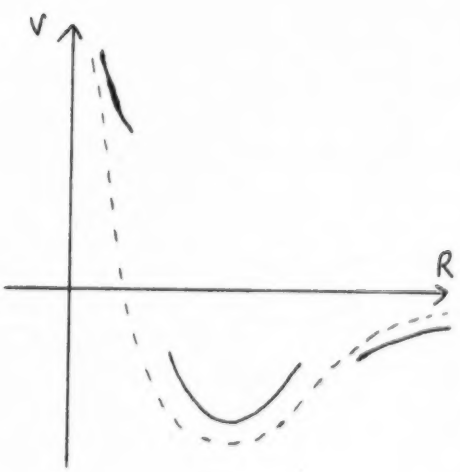


Fig. 2

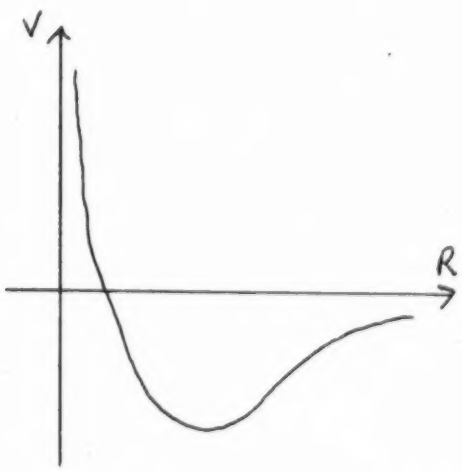


Fig. 3

In case 2 the curvefitted values are first written to the user work area of the disc where the stored values

are already, then all the values are transferred to the graph area. In case 3 the curvefitted values overwrite the stored values and are then transferred to the graph area.

Whatever his reply the user may return to # CHAT to start a new search.

3.3.8 Program DEFL

To obtain a deflection angle it is necessary to have a smooth curve for the potential over the whole range. The potential can be defined by four parameters determined in # CURV; these parameters are used in an equation to give a value for the potential at any point.

The user is first asked for the energy he wants. From this it can be determined if there is no orbiting and the critical impact parameter can be found. These results are given to the user and he is then asked for his choice of impact parameter. The deflection angle is then found and output to the user. He may vary his energy and impact parameter value at will. A particular use of this facility is illustrated in the next section.

3.3.9 Further manipulation

With a smooth representation over the whole range we can obtain quantities like the transport properties which are obtained from the collision integrals (see "Molecular Theory of Gases and Liquids" by Hirschfelder, Curtiss and Bird).

$$\Omega^{(l,s)}(T) = \frac{1}{2}\left(\frac{kT}{2\pi\mu}\right)^{\frac{1}{2}} \int_0^\infty e^{-x}\, x^{s+1}\, Q_l(kTx)\, dx$$

where T = temperature
μ = reduced mass of the two interacting systems
k = Boltzmann's constant.

The collision crossection $Q_l(E)$ depends on the initial relative energy E and is given by

$$Q_l(E) = 2\pi \int_0^\infty b\,(1-\cos^l\chi)\, db$$

where b is the impact parameter and χ is the classical deflection angle

$$\chi(b,E) = \pi - 2b\int_{r_m}^\infty \frac{dr}{r^2\left[F(r,b,E)\right]^{\frac{1}{2}}}$$

in which r_m is the outermost zero of

$$F(r,b,E) = 1 - \frac{V(r)}{E} - \frac{b^2}{r^2}$$

At present the system includes the program to calculate deflection angles, with the user simply providing values for E and b. The next step will be to include

the program to calculate the cross-sections for values of E, and then the collision integral program. These two programs will have to be run off-line, the data being set up on-line.

Note: Further numerical methods are discussed in The Journal of Chemical Physics, 41, pp.3560-3568, (1964).

3.4 A Typical User-System Conversation

We give here a listing of a conversation between an experienced user and the retrieval and manipulation system. The printout is exactly as obtained on a teletype. Lines such as

CHAT: ATOM PAIR?

or

MCS: CHAT (CORE: 13312)

show output to the user from the programs CHAT and MCS (the time-sharing control program supervising the execution of CHAT) while a line such as

: K K

shows information input by the user.

At a number of points we have added brief explanations of the user's replies. These are given at the righthand side in brackets opposite the relevant reply. In particular in these explanations we refer to graphs which are obtained from the user's choices. These graphs numbered one to four are given immediately at the end of this section.

Conversation:

```
    :  £RUN,CHAT
MCS :  CHAT (CORE: 11136)
CHAT:  INTERATOMIC POTENTIALS
CHAT:  FAMILIAR WITH SYSTEM? - "YES" OR "NO"?
    :  YES
CHAT:  ATOM PAIR?
    :  LI K X
CHAT:  ENERGY AND LENGTH UNITS?
    :  AU AU
CHAT:  TYPE "£RUN,FITS"                    [System comments like
    :  £RUN,FITS                            this will be eliminated
MCS :  FITS (CORE: 14336)                   shortly]
FITS:  ALL OR BEST?
    :  ALL                                 [User is asking for all
                                            potentials stored for
FITS:  POTENTIALS STORED FOR LI K X         this state]
FITS:  NO      TYPE                 RANGE (AU)
FITS:  1  VAN DER WAAL'S COEFF  10.9 to 99.0  (LONGRANGE)
FITS:        (97)
FITS:  2   BORN-MAYER             1.5 TO 3.0  (SHORT)
FITS:  99 OTHER FITS
FITS:  WHICH?
    :  1
FITS:  ERROR (RELATIVE):-  10.0%
FITS:  LIST, STORE OR GRAPH?
    :  1 0 1                               [See Graph 1]
FITS:  SHORT DESCRIPTION?
    :  YES
FITS:    V = -C6/R**6
FITS:  WITH R IN AU, C6=0.2290E 04 GIVES V IN AU
FITS:  NO. OF POINTS AND RANGE?
    :  14 20.4 31.6
FITS:  R(AU)              V(AU)
FITS:  21.15          -0.2561E-04
FITS:  21.18          -0.2080E-04
FITS:  22.64          -0.1701E-04
FITS:  23.39          -0.1400E-04
FITS:  24.13          -0.1159E-04
```

```
FITS:  24.88              -0.9655E-05
FITS:  25.63              -0.8085E-05
FITS:  26.37              -0.6805E-05
FITS:  27.12              -0.5756E-05
FITS:  27.87              -0.4890E-05
FITS:  28.61              -0.4173E-05
FITS:  29.36              -0.3575E-05
FITS:  30.11              -0.3075E-05
FITS:  30.85              -0.2655E-05
FITS:  MORE POTENTIALS?
    :  NO
FITS:  NAME,ADDRESS ON 2 LINES
    :  ROBIN MCDONOUGH
    :  78 MALONE ROAD
FITS:  TYPE "ERUN,CHAT"
    :  ERUN,CHAT
MCS :  CHAT (CORE: 11136)
CHAT:  MORE STATES?
    :  A                                    User wants to start
                                              a new search
CHAT:  ATOM PAIR?
    :  K K
CHAT:  STATES STORED FOR K - K
CHAT:  NO.     STATE              NO. OF POTENTIALS
CHAT:  1     K2 X 1 SIGMA+(G)           4
CHAT:  2     K2 B 1 PI(U)               1
CHAT:  99    ANY OTHER STATE
CHAT:  WHICH?
    :  1
CHAT:  ENERGY AND LENGTH UNITS?
    :  AU  AU
CHAT:  TYPE "ERUN,FITS"
    :  ERUN,FITS
MCS :  FITS (CORE: 14336)
FITS:  ALL OR BEST?
    :  ALL
FITS:  POTENTIALS STORED FOR K2 X 1 SIGMA+(G)
FITS:  NO.          TYPE                    RANGE (AU)
FITS:  1     VAN DER WAAL'S COEFF    12.0 to 99.0 (LONGRANGE)
FITS:              123
FITS:  2      RKR                    6.04 to 9.58 (INTERMEDIATE)
```

```
FITS:          (55)
FITS:  3  RKR                  6.24  to  9.15 (INTERMEDIATE)
FITS:     (1)
FITS:  4  BORN-MAYER           1.5   to  3.5  (SHORT)
FITS:  99 OTHER FITS
FITS:  WHICH?
    :  1
FITS:  ERROR(RELATIVE):-            10.00%
FITS:  LIST,STORE OR GRAPH
       0   1   0                        User is asking for his
                                        particular potential to
                                              be stored
FITS:  NO. OF POINTS AND RANGE?
    :  10   12.0   14.0
FITS:  MORE POTENTIALS?
    :  2
FITS:  ERROR(RELATIVE):-            1.0 %
FITS:  LIST,STORE OR GRAPH?
    :  0   1   0
FITS:  MORE POTENTIALS?
    :  NO
FITS:  TYPE  "£RUN,CURV"
    :  £RUN,CURV
MCS :  CURV(CORE:  10880)
CURV:  LIST OR DEFLECTION ANGLE?
       1  0                           User is asking for list of
                                       points fitted to potentials
                                              chosen by him
CURV:  NO. OF POINTS AND RANGE?
    :  20   4.8   12.6
CURV:  FIT FOR K2   X   1   SIGMA+(G)
CURV:  R(AU)                  V(AU)
CURV:  4.8                 0.1008E     00
CURV:  5.21                0.4421E     -01
CURV:  5.621               0.1526E     -01
CURV:  6.032               0.1065E     -02
CURV:  6.442              -0.5245E     -02
CURV:  6.853              -0.7428E     -02
CURV:  7.263              -0.7564E     -02
```

```
CURV:   7.674                -0.6790E   -02
CURV:   8.084                -0.5707E   -02
CURV:   8.495                -0.4613E   -02
CURV:   8.905                -0.3643E   -02
CURV:   9.316                -0.2840E   -02
CURV:   9.726                -0.2204E   -02
CURV:  10.14                 -0.1718E   -02
CURV:  10.55                 -0.1355E   -02
CURV:  10.96                 -0.1091E   -02
CURV:  11.37                 -0.9016E   -03
CURV:  11.78                 -0.7701E   -03
CURV:  12.19                 -0.6807E   -03
CURV:  12.60                 -0.6221E   -03
CURV:  GRAPH?
       I                       User is asking for graph both
                               of the chosen potentials and the
                               fitted values
CURV:  NAME,ADDRESS ON 2 LINES
     : JAMES McLEAN
     : O.U.B.
CURV:  TYPE "£RUN,CHAT"
     : £RUN,CHAT
MCS  : CHAT(CORE:11136)
CHAT:  MORE STATES?
     : I
CHAT:  ENERGY AND LENGTH UNITS?
     : AU   AU
CHAT:  TYPE "£RUN,FITS"
     : £RUN,FITS
MCS  : FITS(CORE:14336)
FITS:  ALL OR BEST?
     : ALL
FITS:  POTENTIALS STORED FOR K2 X I SIGMA + (G)
FITS:  NO.        TYPE                    RANGE
FITS:  I       VAN DER WAAL'S COEFF    12.0   TO   99.0
                                       (LONG RANGE)
FITS:            (123)
FITS:  2         RKR                   6.04   TO  9.58
                                       (INTERMEDIATE)
```

```
FITS:          55
FITS:  3       RKR                          6.24   TO   9.15
                                                (INTERMEDIATE)
FITS:          (I)
FITS:  4       BORN-MAYER                   1.5  TO  3.5 (SHORT)
FITS:  99 OTHER FITS
FITS:  WHICH?
    :  1
FITS:  ERROR(RELATIVE):-                    10.00%
FITS:  LIST,STORE OR GRAPH?
    :  1   1   0
FITS:  SHORT DESCRIPTION?
    :  YES
FITS:  V =  -C6/R**6
FITS:  WITH R IN AU,C6 = 0.3820E 04 GIVES V IN AU
FITS:  NO. OF POINTS AND RANGE
    :  10   12.0   13.0
FITS:  R(AU                      V(AU)
FITS:  12.09                  -0.1223E -02
FITS:  12.18                  -0.1169E -02
FITS:  12.27                  -0.1118E -02
FITS:  12.36                  -0.1070E -02
FITS:  12.45                  -0.1024E -02
FITS:  12.55                  -0.9798E -03
FITS:  12.64                  -0.9383E -03
FITS:  12.73                  -0.8988E -03
FITS:  12.82                  -0.8612E -03
FITS:  12.91                  -0.8254E -03
FITS:  MORE POTENTIALS?
    :  NO
FITS:  TYPE "£RUN,CURV"
    :  £RUN,CURV
MCS :  CURV(CORE: 10880)
CURV:  LIST OR DEFLECTION ANGLE?
    :  0   0                          User wants graph without
                                         examining points.
CURV:  NO. OF POINTS AND RANGE
    :  40  6.0  13.0
CURV:  GRAPH?
    :  1
```

```
CURV:  NAME, ADDRESS ON 2 LINES
     :  L.D.HIGGINS
     :  COMPUTER CENTRE
CURV:  TYPE "£RUN,CHAT"
     :  £RUN,CHAT
 MCS:  CHAT (CORE: 11136)
CHAT:  MORE STATES?
     :  1
CHAT:  ENERGY AND LENGTH UNITS?
     :  AU AU
CHAT:  TYPE "£RUN,FITS"
     :  £RUN,FITS
 MCS:  FITS (CORE:  14336)
FITS:  ALL OR BEST?
     :  BEST
FITS:  TYPE "£RUN,CURV"
     :  £RUN,CURV
 MCS:  CURV (CORE:  10880)
CURV:  LIST OR DEFLECTION ANGLE?
     :  1  0
CURV:  NO. OF POINTS AND RANGE?
     :  16  4.0  9.0
CURV:  FIT FOR K2  X  1  SIGMA+(G)
CURV:  R(AU)                V(AU)
CURV:  4.000              0.2666E 00
CURV:  4.333              0.1751E 00
CURV:  4.667              0.1138E 00
CURV:  5.000              0.7218E-01
CURV:  5.333              0.4363E-01
CURV:  5.667              0.2403E-01
CURV:  6.000              0.1060E-01
CURV:  6.333              0.1485E-02
CURV:  6.667             -0.4613E-02
CURV:  7.000             -0.8596E-02
CURV:  7.333             -0.1107E-01
CURV:  7.667             -0.1250E-01
CURV:  8.000             -0.1320E-01
CURV:  8.333             -0.1340E-01
CURV:  8.667             -0.1325E-01
CURV:  9.000             -0.1286E-01
```

User wants system to obtain best fit to all stored potentials

```
CURV:  GRAPH?
     :  2
CURV:  NAME, ADDRESS ON 2 LINES
     :  GERRY MCGLINCHEY
     :  ANDERSONSTOWN
CURV:  TYPE "£RUN,CHAT"
     :  £RUN,CHAT
 MCS:  CHAT (CORE:  11136)
CHAT:  MORE STATES?
     :  A
CHAT:  ATOM PAIR?
     :  K  KR  X
CHAT:  ENERGY AND LENGTH UNITS?
     :  AU  AU
CHAT:  TYPE "£RUN,FITS"
     :  £RUN,FITS
 MCS:  FITS (CORE:  14336)
FITS:  ALL OR BEST?
     :  BEST
FITS:  TYPE "£RUN,CURV"
     :  £RUN,CURV
 MCS:  CURV (CORE:  10880)
CURV:  LIST OR DEFLECTION ANGLE?
     :  0  1
CURV:  TYPE "£RUN,DEFL"
     :  £RUN,DEFL
 MCS:  DEFL (CORE:  9216)
DEFL:  UNITS TO BE USED:
DEFL:  LENGTH IN AU AND ENERGY IN AU
DEFL:  REQUIRED ENERGY, E?
     :  1.234
DEFL:  ORBITING OCCURS IN THE RANGE 1.14005 (AU) TO
DEFL:  1.32153 (AU)
DEFL:  CRITICAL IMPACT PARAMETER = 0.336 (AU)
DEFL:  IMPACT PARAMETER, b?
     :  0.9464
DEFL:  FOR E=1.234 (AU) AND B=0.946 (AU),
DEFL:  DEFLECTION ANGLE = 0.1425E 01 (RADIANS)
```

User wants graph of fitted potential alone. See Graph 4.

```
DEFL:  ANOTHER IMPACT PARAMETER, B?
    :  2.865
DEFL:  FOR E=1.234 (AU) AND B=2.865 (AU),
DEFL:  DEFLECTION ANGLE = 0.6290E 00 (RADIANS)
DEFL:  ANOTHER IMPACT PARAMETER, B?
    :  3.498
DEFL:  FOR E=1.234 (AU) AND B=3.498 (AU),
DEFL:  DEFLECTION ANGLE = 0.4632E 00 (RADIANS)
DEFL:  ANOTHER IMPACT PARAMETER, B?
    :  NO
DEFL:  A DIFFERENT ENERGY, E?
    :  1.185
DEFL:  ORBITING OCCURS IN THE RANGE 0.114005 (AU) TO
DEFL:  1.32153 (AU)
DEFL:  IMPACT PARAMETER, B?
    :  0.876
DEFL:  FOR E=1.185 (AU) AND B=0.876 (AU),
DEFL:  DEFLECTION ANGLE = 0.1539E 01 (RADIANS)
DEFL:  ANOTHER IMPACT PARAMETER, B?
    :  5.803
DEFL:  FOR E=1.185 (AU) AND B=5.803 (AU),
DEFL:  DEFLECTION ANGLE = 0.1149E 00 (RADIANS)
```

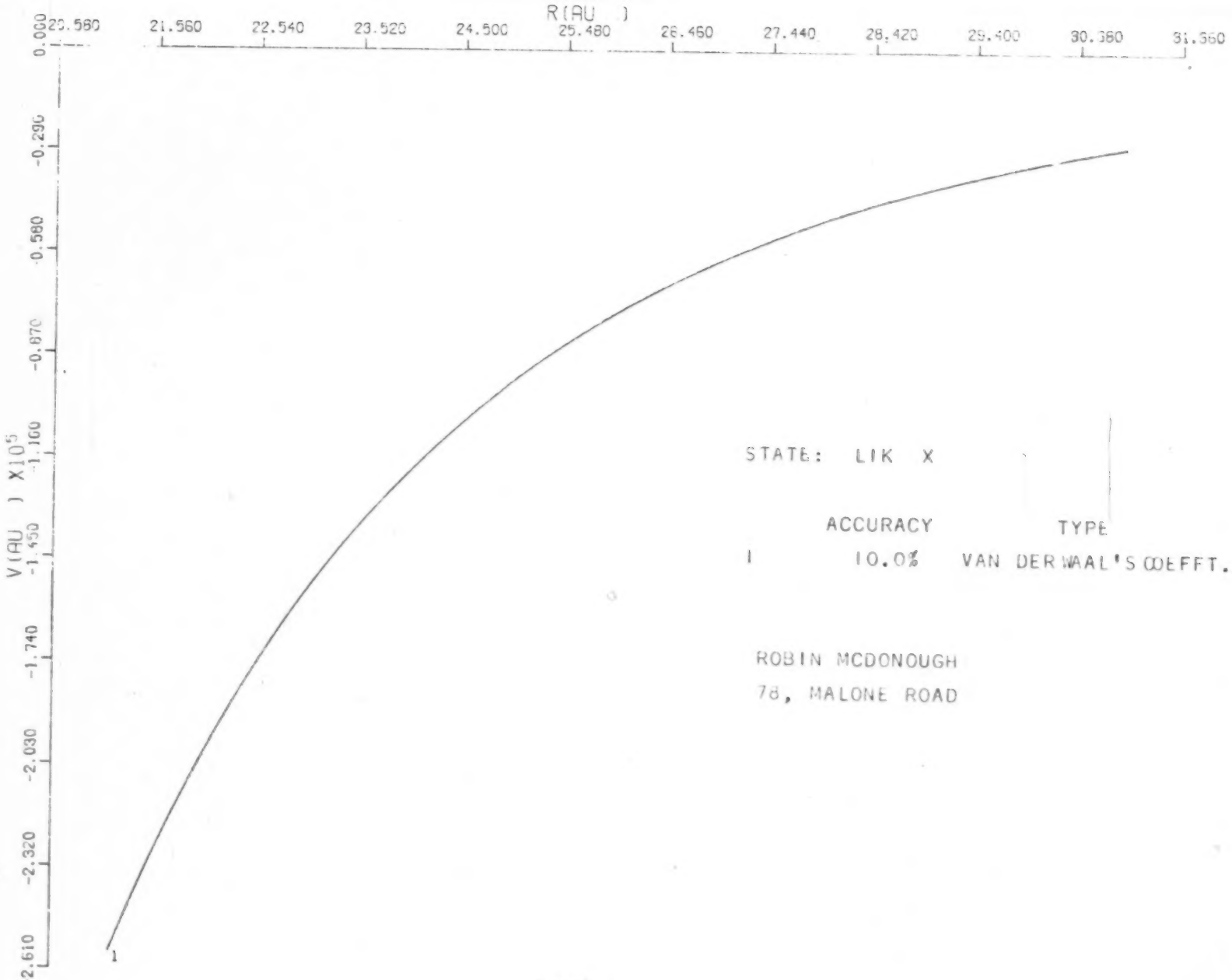
* * * * *

The next search is for a different potential

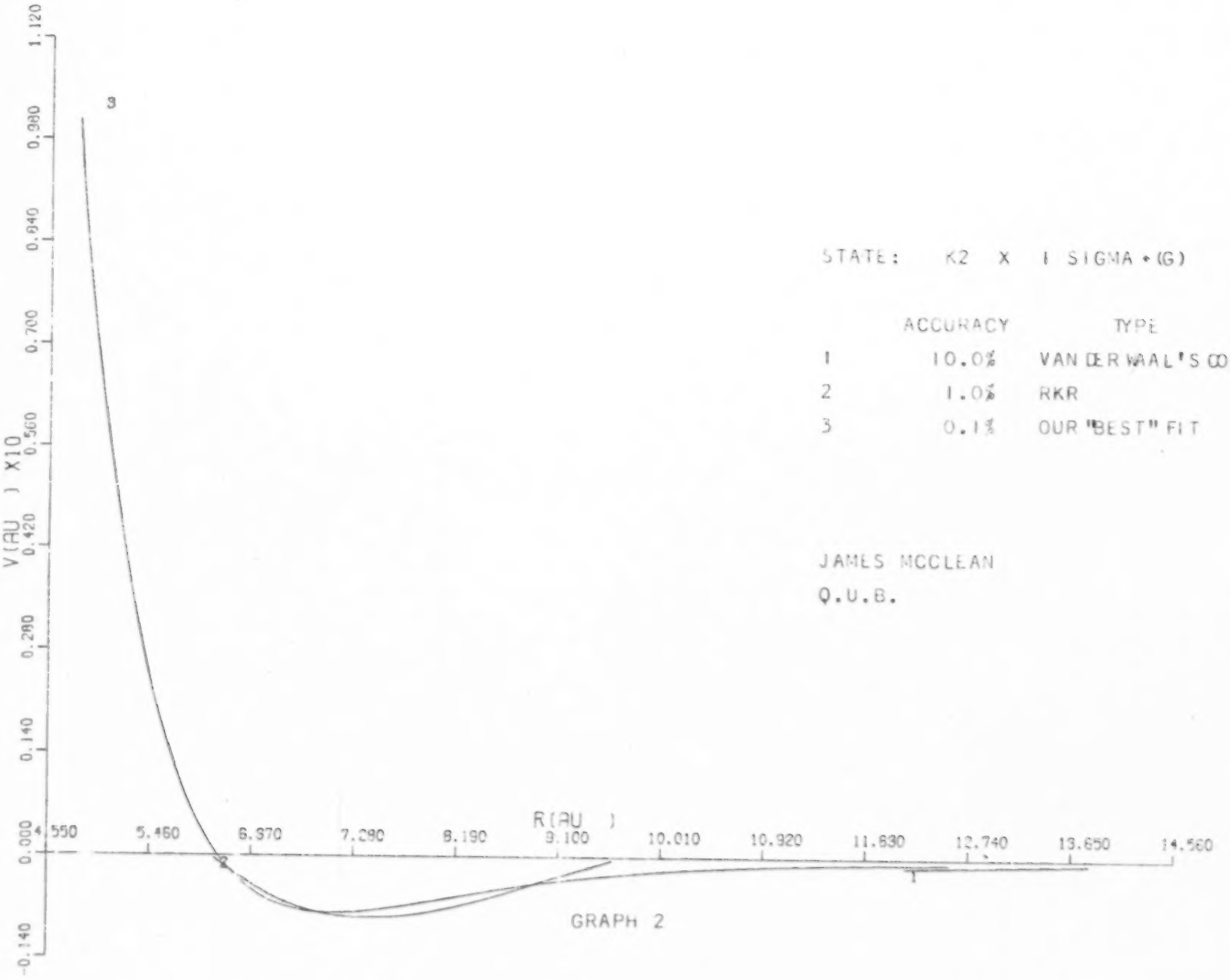
* * * * *

```
DEFL:  REQUIRED ENERGY, E?
    :  0.3825
DEFL:    "NO ORBITING"
DEFL:  CRITICAL IMPACT PARAMETER = 0.000 (AU)
DEFL:  IMPACT PARAMETER, B?
    :  1.007
DEFL:  FOR E=0.3825 (AU) AND B=1.007 (AU),
DEFL:  DEFLECTION ANGLE = 0.3143E-08 (RADIANS)
DEFL:  ANOTHER IMPACT PARAMETER, B?
    :  NO
DEFL:  A DIFFERENT ENERGY, E?
    :  16.14
```

```
DEFL:     "NO ORBITING"
DEFL:  CRITICAL IMPACT PARAMETER = 0.0000E 00 (AU)
DEFL:  IMPACT PARAMETER, B?
    :  0.1716
DEFL:  FOR E=16.140 (AU) AND B=0.1716 (AU),
DEFL:  DEFLECTION ANGLE = -0.8149E-09 (RADIANS)
DEFL:  ANOTHER IMPACT PARAMETER, B?
    :  NO
DEFL:  A DIFFERENT ENERGY, E?
    :  NO
DEFL:  TYPE "ERUN,CHAT"
```


R(AU)
20.580
21.560
22.540
23.520
24.500
25.480
26.460
27.440
28.420
29.400
30.380
31.360
V(AU) X10^5
0.000
-0.290
-0.580
-0.870
-1.160
-1.450
-1.740
-2.030
-2.320
-2.610
STATE: LIK X
ACCURACY
TYPE
1
10.0%
VAN DER WAAL'S COEFFT.
ROBIN MCDONOUGH
78, MALONE ROAD

GRAPH 1

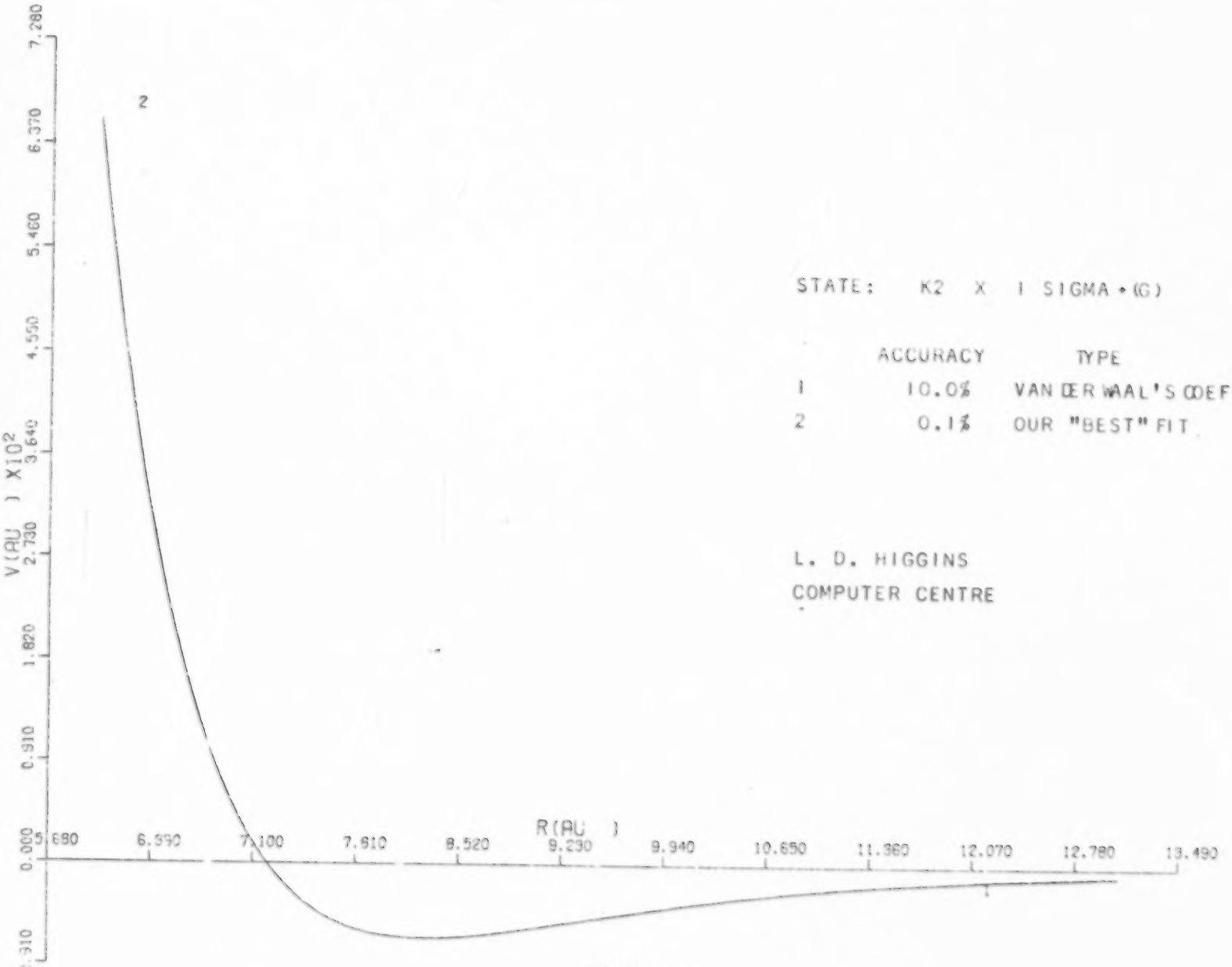


GRAPH 2

STATE: K2 X 1 SIGMA+(G)

	ACCURACY	TYPE
1	10.0%	VAN DER WAAL'S COEFFT.
2	1.0%	RKR
3	0.1%	OUR "BEST" FIT

JAMES MCCLEAN
Q.U.B.

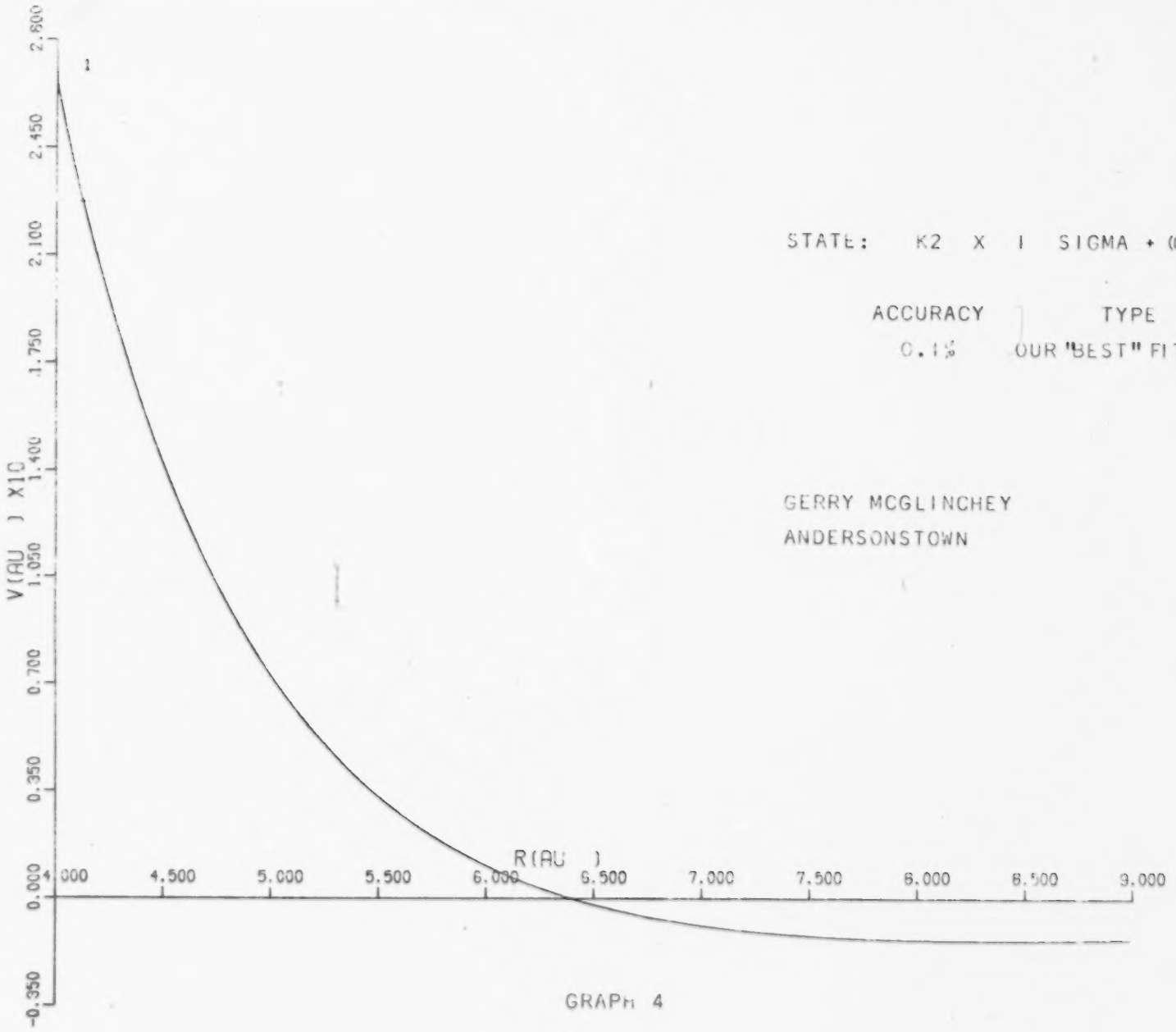


GRAPH 3

STATE: K2 X 1 SIGMA + (G)

	ACCURACY	TYPE
1	10.0%	VAN DER WAAL'S COEFFT.
2	0.1%	OUR "BEST" FIT

L. D. HIGGINS
COMPUTER CENTRE


STATE: K2 X 1 SIGMA + (G)
ACCURACY
0.1%
TYPE
OUR "BEST" FIT
GERRY MCGLINCHEY
ANDERSONSTOWN
V(AU) X10
R(AU)
2.800
2.450
2.100
1.750
1.400
1.050
0.700
0.350
0.000
-0.350
4.000
4.500
5.000
5.500
6.000
6.500
7.000
7.500
8.000
8.500
9.000

GRAPH 4

3.5 Miscellaneous Developments

3.5.1 Editing program

A program named C124 has been written to enable the potential data once stored to be altered. It is envisaged that this program will be needed in three different situations.

Firstly, alterations may be required to correct mistakes inadvertently made when the data was first stored. The program will in particular have to allow for changes in potential values as well as in titles, references, and the various codes used to describe the type and range of the potential. Secondly, the assessment of the accuracy and range of validity of a particular potential by a consultant expert will in many cases differ from our original evaluation and so the capability to effect this change will be needed. Finally, it may be necessary from time to time to replace some information already stored. For example, more accurate potential well depths may become available, or additional relevant information might be added to that already stored. The present program allows for the possibility of all these changes.

The actual editing is carried out on-line and a number of examples of how this is done are given at the end of this section.

3.5.2 Checking program D123

This program writes out the list of potentials of the same type for pairs of atoms. It enables an easy comparison to be made of all the values stored for a particular potential describing a particular diatomic interaction.

The use made of this program will be twofold. Firstly, by comparing one stored potential for a system with another for the same system, a check for possible errors in extraction and storing can be made. Secondly, it will present the potential data in a form which can be easily evaluated by our outside consultant experts.

Experience has shown the necessity of a careful check on the data stored and with the database of potentials now almost complete this can now be carried out with the use of this program.

Example of Numerical Data Editing Program C124

This program alters or deletes complete potentials. Complete potentials refer to statelist, reference record and potential record. If more than one section is being changed at any time, they must be taken in the following order:-

1. Statelist
2. Reference Record
3. Potential Record

The potential record is made up of five small sections and the program considers each one in turn and makes the appropriate alterations. If the potential record has been found to be useless, then the complete potential is eliminated.

We here give a typical conversation between the user and the program C124 when we, for example, make the following changes:-

1. Eliminate a potential
2. Change a state name
3. Change a reference
4. Change an actual potential value.

Lines such as

C124: POTENTIAL ADDRESS

or

MCS: C124 (CORE: 12032)

show the output on the teletype to the user from the programs C124 and MCS (the time-sharing control program supervising the execution of C124) respectively, while a line like

: 51947

shows information input by the user.

```
    :  £RUN,C124
 MCS:  C124 (CORE: 12032)
C124:  POTENTIAL ADDRESS?
    :  51947
C124:  ELIMINATE A POTENTIAL?
    :  YES
C124:  (1) BUCKET NO OF STATE?  (2) STATE NO?  (3) PREVIOUS
       POTENTIAL ADDRESS?  (THIS IS ZERO IF FIRST POTENTIAL
       IN STRING IS BEING ELIMINATED)  (4) NEXT POTENTIAL
       ADDRESS?  (THIS IS ZERO IF LAST POTENTIAL IN STRING
       IS BEING ELIMINATED)
    :  98    1    51895    52575
C124:  CHANGE MORE POTENTIALS?
    :  YES
C124:  POTENTIAL ADDRESS?
    :  49073
C124:  ELIMINATE A POTENTIAL?
    :  NO
C124:  WHICH TYPE OF ALTERATION?  GIVE NUMBER ONLY (CHANGES
       TO EACH COMPLETE POTENTIAL MUST BE IN THE ORDER
       GIVEN)  (1) STATELIST   (2) REFERENCE RECORD
       (3) POTENTIAL RECORD
    :  1
C124:  BUCKET NO?   STATE NO?
    :  127    1
C124:  NEW STATE NAME?
    :  HG   HE   X
C124:  HG   HE   X
C124:  MORE ALTERATIONS TO SAME POTENTIAL?
    :  YES
C124:  WHICH TYPE OF ALTERATION?  GIVE NUMBER ONLY
    :  2
C124:  TITLE OR ARTICLE?
    :  LONG-RANGE INTERACTIONS OF MERCURY ATOMS
C124:  LONG-RANGE INTERACTIONS OF MERCURY ATOMS
C124:  AUTHOR NAME(S)?  JOURNAL NAME?  ETC.
    :  W.C. STWALLEY & H.L. KRAMER    J. CHEM. PHYS. 49,
       5555 (1968)
C124:  MORE ALTERATIONS TO SAME POTENTIAL?
    :  YES
```

```
C124:  WHICH TYPE OF ALTERATION?  GIVE NUMBER ONLY
    :  3
C124:  CHANGE ACTUAL POTENTIAL VALUES?
    :  YES
C124:  NEW VALUE?
    :  15.30
C124:  15.30
C124:  CHANGE MORE POTENTIALS?
    :  NO
C124:  HALTED:- 00
    :  £ENDJOB.
```

4. ASSESSMENT OF PRESENT POSITION

In assessing the present position of the numerical databank we looked separately at the three branches of the work, viz. data extraction, potential representation and retrieval and manipulation.

With the extraction of the interatomic potentials now almost complete there are three separate matters to be dealt with in the future. Firstly, the data base of interatomic potentials once completed must be kept up to date with all new publications. Secondly, a check of the stored potentials must be made both to ensure no errors have been made in storing and also to complete, with the help of our outside consultants, the evaluation of the data. Finally the subsidiary data bases of quantities such as oscillator strengths and polarizabilities must be created.

The problem of obtaining a satisfactory representation of the potential energies to enable manipulative procedures to be readily applied to them continues to be the most difficult one facing us. Our intention is to continue to seek a form of the potential data to which polynomials can be best fitted. Section 3.2 gave some indication of the "reduced" form of potential curves which have so far been tried. Should no such suitable "reduced" form be obtained a straightforward interpolation procedure using the stored potential points will be used as an alternative.

Finally, the retrieval and manipulation side of the system in its present form not only allows for the retrieval of the stored potentials in tabular sets of chosen co-ordinates and/or their respective off-line graphical representations but has made a start towards the inclusion of programs for deriving useful results from these potentials. To date, part of a large program for calculating transport properties of gases to any accuracy has been implemented into the system; this allows for the on-line calculation of deflection angles (section 3.3). However, whilst the deflection angle procedure and other identifiable parts of the intended large transport

property program can be implemented in an on-line mode, it is not practical to aim at including the complete program in this way; instead we aim to be able to initiate it via a remote job entry, similar to the graphplotting program (section 3.3).

We hope further to encourage the future use of our system by allowing users not just to avail themselves of the fixed number of operations that we have incorporated but also to be able to extract and, if necessary, manipulate with our data up to the point where they can then store it into their own personal files and apply their own suite of programs to it.

In conclusion, a significant new dimension to the system is beginning to blossom. No longer do we think of our system just as a tool for storing numbers, rather we believe the true value of the system lies in the dynamic concepts of allowing the user to apply his own initiative in obtaining further data. The desirability of such features can only be judged by experience.

Appendix A1

≠ QUIS PROJECTS

L. D. HIGGINS, The Queen's University of Belfast, N. Ireland.

1. INTRODUCTION

Two main on-line systems have been developing at Queen's University, Belfast, for nearly five years: a reference retrieval system and a numerical data system. Rather than provide a current awareness or retrospective service and so, cater for static type profiles, we aim at offering a flexible service to the user who is prepared to build up his own profile interactively in his own time. I will first outline our reference retrieval project and then the numerical data retrieval project.

2. QUOBIRD REFERENCE PROJECT

We began by first scrutinizing the literature, but even as recently as four or five years ago, there did not appear to be much of value to be found. We then decided to develop our own system; in this way we hoped to discover the problems involved for ourselves as our system evolved. The first real result of our efforts was the implementation of an on-line experimental system, which has been operational for over three years, and on which we have been developing ever since.

The essence of the indexing side of the system consists of a subject-inverted-title and abstract-file. Although we retrieve references for a variety of scientific subjects our main data base today consists of approximately 3,500 atomic and molecular physics abstracts along with their associated references. These abstracts are extracted directly from bi-monthly Inspec tapes, which we have been receiving for the past 18 months or so. Prior to this our content matter was taken from scientific books, namely their titles and chapter headings together with the bibliographic details of each book. We first indexed our local departmental computer science library and since then we have indexed similar libraries in physics and applied mathematics. Alongside the indexing of the latter libraries we began extracting papers taken from Inspec tapes; treating the sentences of the abstracts belonging to the papers in a similar manner to that already adopted for the capter headings of the books. In this way the basic indexing design needed little change. A list of the Inspec tape subject headings which we use is given in Figure 1.

13.00	Atomic and Molecular Physics
13.20	Atoms
13.23	Hydrogen and Helium Atoms
13.25	Isotopes
13.30	Molecules
13.31	Inorganic Molecules
13.37	Intermolecular Mechanics

FIGURE 1. Inspec Headings for Physics Classification

We have in a special report, SR6, an up-to-date alphabetical print-out of the dictionary as exists in our atomic and molecular physics data base file at present; along with some additional statistical information.

We have attempted to implement our experimental operational pilot system under the following design criteria:

(a) Self-instructive
(b) Simple to use
(c) Rapid Response
(d) Effective to use
(e) Minimum Cost
(f) Efficiency of Storage
(g) Minimisation of disc accesses

FIGURE 2.

For detailed descriptions of these headings I refer you elsewhere (Refs. 1 & 2). It is my intention to outline the facilities offered by the system and to indicate how users of it are influencing the on-line conversational searching language. The best way to do this is to show you a brief illustration (Figure 3) of what a user might experience during a typical search. I have chosen as a data base the atomic and molecular physics abstracts extracted from the Inspec tapes and stored on our fixed disc store. Although the system is aimed at being self-explanatory, I will first outline the facilities that are offered.

The title and abstract records (or title and chapter headings records) can be retrieved using a key phrase consisting of up to 8 words and the possible alternatives offered are:

(a) a further modification of the user's key phrase to expand or contract the set of documents retrieved;

(b) the set of sentences (or chapter headings) containing the keywords;

(c) the complete abstract of the documents;

(d) the citations of the documents, i.e. the title, author, publisher, volume and issue number.

These points can be seen in the following example:

≠ Paper read at NATO Summer School on On-line Mechanised Information Retrieval Systems.

```
BIRD :    KEY?
     :    CROSS SECTION
BIRD :    KEY: 107 DOCS/ 206 HDS
BIRD :    R, H, T, OR M?
     :    M
BIRD :    KEY?
     :    HYDROGEN ATOM
BIRD!:    NEW KEY: 27 DOCS/43 HDS
BIRD :    INTERSECTION: 8 DOCS/ 21 HDS
BIRD :    UNION: 126 DOCS/ 244 HDS
BIRD::    N, I, P OR U?
     :    I
BIRD :    R, H, T OR M?
     :    M
BIRD :    KEY?
     :    PROTON IMPACT
     :
     :
     :    ELECTRON CAPTURE
     :
     :
BIRD :    N, I, P OR U?
     :    I
BIRD :    R, H, T OR M?
     :    T
BIRD :    PRINT OUT OF ABSTRACT
```

(The options were:
(R: references
(H: headings
(T: text of abstract
(M: more keys

(The options were:
(N: documents of new key
(I: intersection of documents
(P: documents for previous key
(U: documents for new and previous key

FIGURE 3. SAMPLE OF BIRD PROGRAM

3. QUODAMP NUMERICAL PROJECT

We turn now to our numerical data on-line system. Shortly after the commencement of the reference retrieval project, we believed that although a paper might be important to a scientist, frequently it is the numerical data given there that is really required. To take this kind of retrieval process to its ultimate goal, we feel that the main facility a numerical on-line data system could offer would be to enable the user not only to retrieve a set of relevant numbers, but to be able to manipulate with these numbers to calculate results he wanted them for in the first place. After four years experience we are confident that this is possible.

When we started this project we decided to store interatomic potentials as numerical data for our research. Since we had an applied mathematics department which specialised in atomic and molecular physics we felt that we might be able to offer some kind of local experimental service to its members who in turn could help us by providing us with critical feedback from the system. Today the beginnings of such a system is operational and has been so for two years. Like the reference retrieval system, the operation of it requires little or no knowledge on the part of the user. He communicates by answering multiple-choice questions in English. Currently the system allows retrieval from a data bank of more than 500 interatomic potentials. The design of this system has been very much dictated by the structure of the data itself, which I will now briefly outline and then mention some of the facilities offered. For more detailed information I refer you to the annual report (Ref. 3) and to Ref. 4.

Each pair of atoms can be in any of an infinite number of states, the unexcited state being classed as the 'ground state'. For each state there is a potential function $V(r)$, $0<r<\infty$ which represents the forces between the two atoms. Approximations to $V(r)$ over different ranges of r are obtained by both theoretical and experimental means. There are many approximations in different ranges of r by different people and to different accuracies. An illustration of what a potential looks like can be seen in Figure 4. In the data bank are stored all the fits to the potentials that can be obtained from the literature; the references are obtained mainly from Physics Abstracts, those from 1965-70 being scanned to date. The fits can be tabular sets of values or coefficients to be inserted in pre-defined formulae. The data is stored in atomic units but it can be retrieved in any desired units from relevant information like accuracy, range of validity and source.

The system in its present form allows for the retrieval of any of the stored potentials. As well as the ranges and accuracies reported the units can be changed to suit the needs of the user. The user can have an off-line graph of the potential if he wants it and/or a pseudoplot of it on-line. Presently programs to derive useful results from the potentials, like the transport properties of gases, are being included. Such programs are already in existence but before they can be used a means has to be found to derive from the different fits a smooth representation of the potential over the whole range $0<r<\infty$. Among other things this curve fitting problem illustrates the kind of problem we are experiencing in this system, but once again we remain confident. An illustration of the pilot on-line system is given in Figure 5.

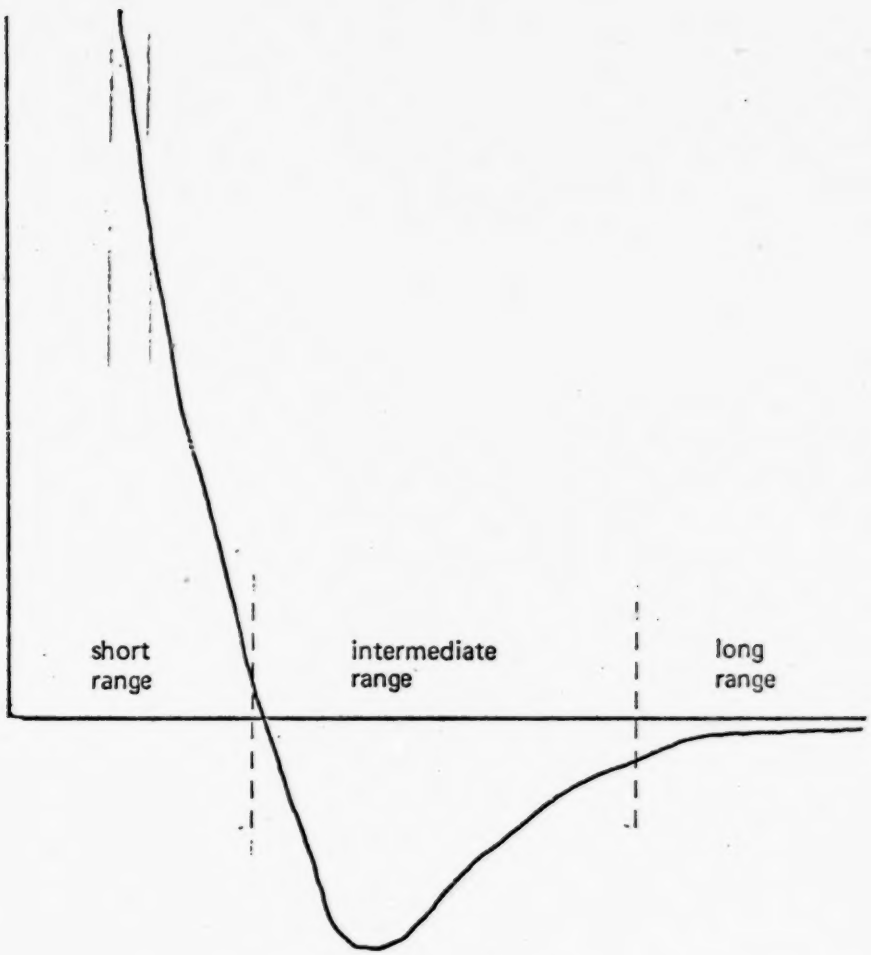


FIGURE 4. Typical shape of an interatomic potential. The dashed lines give a rough idea of the boundaries between the short, intermediate, and long range regions of the potential.

```
DAMP  :   TYPE YOUR TWO ATOMS
      :   LI H
DAMP  :   LIST OF STATES STORED FOR LI – H
DAMP  :   1  LIH  X  1  SIGMA
DAMP  :   2  LIH  B  1  PI
DAMP  :   3  LIH  A  1  SIGMA +
DAMP  :   99 ANY OTHER STATE
DAMP  :   WHICH?
      :   2
DAMP  :   ENERGY AND LENGTH UNITS?
      :   A.U.  A.U.
DAMP  :   OPTIONS?
      :   ?
DAMP  :   THE OPTIONS ARE
DAMP  :   1  PARTICULAR FIT(S)
DAMP  :   2  BEST POTENTIAL
DAMP  :   3  ALL THE FITS STORED
DAMP  :   4  VALUES IN A CHOSEN RANGE
      :   2
DAMP  :   GRAPH?
      :   NO
DAMP  :   NO. OF POINTS AND RANGE?
      :   10  1.0  4.0
DAMP  :   VALUES OF BEST FIT
DAMP  :   R(A.U.)   V(A.U.)
DAMP  :   0.1000E 01   0.9647E 00
                :          :
                :          :
DAMP  :   PSEUDOPLOT?
      :   YES
DAMP  :   0.9647E 00 ...
          0.5127E 00 ...
          0.6070E-01 ...
DAMP  :   MANIPULATE?
      :   NO
```

FIGURE 5. Example of DAMP Program

In conclusion, the data system is not in as advanced stage of development as the reference system which has had several users, both local and abroad. This experience has helped us realise the type of user language that is needed, and we now believe that our indexing approach has to be greatly improved.

Acknowledgements

This work was supported in part by the Office of Scientific and Technical Information, London.

References

1. L. D. Higgins and F.J. Smith, *Computer Journal, Vol. 14,* (1971), pp. 249-53.
2. M. Carville, L.D. Higgins and F. J. Smith, *Information Storage and Retrieval,* Vol. 7, (1971), pp. 205-210
3. Queen's University On-Line Data Bank on Atomic and Molecular Physics, *Annual Activity Report, 1971.*
4. F. J. Smith, Physics of Electronics and Atomic Collisions, *VII ICPEAC, 1971* - North Holland (1972)

Appendix A2

REPORT TO OSTI ON THE NATO SUMMER SCHOOL ON ON-LINE INFORMATION RETRIEVAL SYSTEMS

attended by

E D Barraclough, Newcastle University, L Higgins, Queens University, Belfast, I McCracken, UKCIS, Nottingham University, C J van Rijsbergen, Cambridge University

The Organisation of the Course

The Summer School was split into two parts; the first week consisted of formal presentations by invited speakers. These were very much a survey of the state of the art of on-line systems at the present time. During this first week very little time was allowed for discussion and only questions of clarifications were answered during the lectures. In practice most of the information presented was not controversial so this did not raise too many problems.

In the second week brief presentations were made by members of the course, generally on work that was actually in progress. The remainder of the week was then spent on panel discussions covering the main areas of the subject. Three of the OSTI supported participants gave talks on the work they were doing and two sat on the panels. In the evenings demonstrations were arranged of the working on-line systems. NLM's Medline was demonstrated on-line to Sweden; the Culham system was demonstrated from a video terminal over dial up line to Culham Laboratory; the Newcastle Medlars system was demonstrated from a 2741 to the computer at Newcastle; and the Datacentralen system was demonstrated to the computer in central Copenhagen. A demonstration of both the Belfast reference retrieval system and numerical date system was also given using the Culham video terminal. It was not possible to link up to the Spires-system at Stamford due to incompatibilities between the modems in Europe and those in the US. The formal part of the course was thus very well organised. However, it is generally the informal part that proves most valuable, and here we suffered as there were no communal eating facilities at the student home where we were staying. As a result the group tended to disperse in the evenings, in particular most of the lecturers disappeared into Copenhagen and were not available for informal discussions.

Proceedings of the Summer School will be published in due course.

The main areas covered

The formal presentations could be divided into a section on the theory behind information retrieval systems which were mainly concerned with clustering techniques; secondly the hardware available in current computing systems and its impact on information retrieval systems. Here the emphasis was on the cost and speeds of storage media and transmission facilities. The third topic was the design and implementation of the systems where the concern was mainly with

the user interface. The last major area was the management and the evaluation of these systems.

Theoretical Aspects

Conventionally the information retrieval literature distinguishes broadly between three types of file organisation: sequential, inverted and clustered. The distinction is often convenient but can be misleading. Each type is in fact a special case of a clustered file-structure.

A clustered file structure consists of two things, a set of clusters and a set of cluster representatives (commonly called centroids) where a cluster representative characterises (summarises) the cluster. A moment's thought will show that an inverted file is a primitive clustering. In fact it is a clustering in which each cluster is represented by one and only one index term. The clusters also overlap to an arbitrary extent. Similarly, a sequential file is an extreme case of a clustered file - each cluster contains only one document and is represented by the index terms of that document. The point here is that a clustered file is not different in kind from an inverted file but in degree.

The debate as to which of these file structures to use for on-line document retrieval has now centered about the inverted versus the clustered file. It is accepted now that for on-line document retrieval sequential files are inadequate. The response time is too slow although the effectiveness may be greater than that achieved with an inverted file. So, a major advantage of the inverted file is its retrieval speed. However, it has been shown that retrieval based on hierarchic document clustering can achieve the effectiveness of a linear search followed by ranking. Potentially cluster-based retrieval is more effective than any ranking method. Intuitively this follows from the fact that clustering brings together documents relevant to the same queries while at the same time separating the relevant from the non-relevant. The experimental results supporting this claim must be viewed with caution since they have only been obtained on relatively small data-bases.

Professor Salton criticised the use of inverted files on a number of grounds. The first of these (in which he was supported by Mr Cleverdon) is that inverted files are nothing more than glorified peek-a-boo systems. The implication being that we are not exploiting computer technology to the fullest extent but are using the same techniques for information retrieval as were used before computers were introduced. The second objection is that inverted files limit one to Boolean searches whereas it has now been established that ranking methods, using sophisticated matching functions, are more effective. Thirdly it is impossible

to implement feedback procedures when operating on an inverted file. Fourthly one is stuck with a static indexing system since it is costly to update an inverted file with respect to indexing. Professor Salton claims that the answer to all these problems lies in automatic document clustering. Unfortunately clustered file structures have only been tried on a relatively small scale, and only in an experimental environment. It would seem that the testing of automatic document clustering on a large data-base is long overdue. It is true that the clustered files require an initial investment of order $n\log n$ to n^2 in CPU time for its construction but then the construction of a flexible inverted file is not cheap either. The extra performance and flexibility achieved from a clustered file would seem worth it.

Nevertheless, it was not universally accepted that clustering was necessarily the best method of structuring files. Some people felt it dangerous to let users have control of the data base structure without completely understanding it. Two large operational systems using inverted file techniques were those illustrated by Professor Parker using the Spires system and by Dr Katter and Mr McCarn using the Medlars system.

Madame Wolff-Terroine gave a survey of some clustering methods. Unfortunately the survey was very sketchy and did not contain any of the theoretical results obtained in the last three to four years. She discussed her use of clustering in keyword classification which was mainly based on the work done by K Sparck Jones. Unfortunately no attempt was made to evaluate the experimental work except by visual inspection of the classifications.

Madame Wolff-Terroine in her presentation also hinted at the difficulty of automatically finding the content units to describe a document. Mr Cleverdon elaborated on this difficulty by stating that it was not possible to consider the specificity of the content units independent of the level of exhaustivity of the document description.

Hardware

Dr Helms from the Computing Centre gave a series of talks on the capabilities of both the hardware and software of present computing systems. He estimated that in Europe we were still two years behind the US. One area where vast improvement could be foreseen is in the provision of large computer stores. For example, a store of 10^{12} bits is quoted as costing $\$10^6$. The problem with information retrieval systems is not only the size of the store required but the data transfer rate between the storage device, the computer and the user. Matching these is the problem of the software designer. Dr Helms quoted some figures giving the times in man-years required to implement operating systems.

For example, IBM's OS system took 5000 man-years of effort to reach its present state. The complexity of such a system can affect its reliability and when one is running an on-line system with many remote users reliability is all important. Unfortunately for designers of on-line retrieval systems it is not possible to control the operating system that is being used. This rather than the information retrieval system itself could well be the major area of difficulty.

Design and Implementation

Dr Katter of Systems Development Corporation described the requirements for the design of an on-line system. He considered this from a commercial point of view in that they were concerned not only to provide a working system that was attractive to the user but also to make such a system economically viable. The main topics that were considered were the maintenance of the data-base, that is how to validate the data being added and how to control the size of the data-base by selecting items to be purged. This last problem really had no satisfactory solution. Also on the control side one required statistics showing the usage of the system. One needed facilities for file security and in a commercial system for accounting and billing. For the user of the system one clearly had to provide searching capability and here the interface with the user was all important. The user also needed the facility to print a sample of the citations on the file and options in the form of the output. Many of the systems being demonstrated showed the facilities that were described.

Management and Evaluation

Two speakers covered these topics: Mr McCarn from the National Library of Medicine was concerned with the management of a large system and Professor Lancaster from University of Illinois talked about evaluation methods. The main problems with management of such systems are in the communications area. With many users spread throughout the United States they could not afford to contact individuals in the case of a breakdown of the system. Nor could they afford to use the normal telephone network for communicating over such vast distances. Both these problems have very little to do with the computer or the software system that is running on it. They are almost entirely a communications problem arising from the fact that data communication of this type takes a much longer time and uses a telephone line very inefficiently compared with normal speech. The solution that they are attempting in the States is the provision of networks of lines controlled by small computers which can pack messages and thus communicate much more efficiently. This also partially solves the problem

of machine breakdown in that the user can get information from the communications network or he can be transferred to a different machine. In the United Kingdom there are no working networks covering a wide area and the Post Office's plans for such a system are very remote. The need for better and cheaper communications is obvious when one considers the costs of running a search on the Medline system. This dropped to as low as $5 if 30 people were using it simultaneously.

Professor Lancaster gave a survey of the various on-line systems available and the requirements from an on-line system. He felt that despite the inaccuracy in the measurements of precision and recall these were the only measurements that could be used for testing systems and maintained that a user on-line could by sampling the file increase his precision and save a lot of computer time by not doing abortive searches. He advocated ranking the output so that the user saw the most specific documents first. This was particularly important in the on-line system where the number of documents a user would see would be relatively small.

Panel Discussions

Four main areas of on-line systems were discussed, first the training of users and here there was some difference of opinion concerning the use of computer aided instruction techniques for this type of system. At present the only people actually training users are Medline where the National Library of Medicine spends three weeks training librarians. Most of this was not spent on the computer system but rather on understanding the Mesh vocabulary and the indexing requirements. The librarians attending the Summer School felt that it should be possible to train for the general use of on-line systems and not for a particular system.

The next panel discussion was concerned with the interface between the user of the system. There seemed to be no panacea and no clear way of distinguishing a good system from a bad system. Devising an experiment to compare two systems would be very difficult.

File organisation was another topic and the only lesson to be learnt was that different types of file are suitable for different file sizes and methods of use. No-one had done any cost analysis relating to file type and size. This was always left to the system implementor and at present the amount of theory is very limited. The final panel discussion was on cooperation between libraries. The situation in the United States seems even less hopeful than it is here, one

of the reasons perhaps being that they have too much money. A plea was put in for the extension of British MARC to cover the European literature as well. Cooperation on on-line systems seems only possible on a cost basis for the libraries in one locality.

The Future of on-line Systems

It was clear by the end of the Summer School that on-line systems where the user interrogated the system himself had come to stay. The most fruitful area for research would seem to be in designing systems for the interrogation of more than one data base. It was also felt that the user did not need to have access to the complete retrospective file. He wanted only a few relevant references to begin with. Retrospective searching could then be done if necessary at a later date on a batch system.

Areas requiring further investigation

It was apparent that there was a need to test the theory of information systems in a real situation. Clustered files are potentially more flexible and effective than inverted files. The evidence for this is pretty slim so far being based only on small files. It is essential that more research is done to prove (or disprove) this claim. One way this can be done is by mounting a large scale automatic clustering experiment.

The user oriented research for which a need became apparent was on the application of on-line systems to more than one data base. There seemed to be two levels that could be distinguished. The actual interrogation of a large data base from a formulated query and the assistance for the user during the formulation process. At present attempts are made to include both facilities in one system with the emphasis on different aspects according to the apparent needs of the users. For example, the Medline system concentrated initially on interrogation of the data base while the Newcastle Medlars system was concerned with user assistance. For large scale systems it is clearly more efficient to keep the tutorial aids to a minimum, thus reducing the message processing requirements.

A method of overcoming this conflict in requirements is to provide tutorial aids on a satellite computer system. Thus, the user would for tutorial purposes interact with the satellite computer, and would only interrogate the central data bank when he had reached a predefined level of proficiency. Figure 1 shows such a dual system:

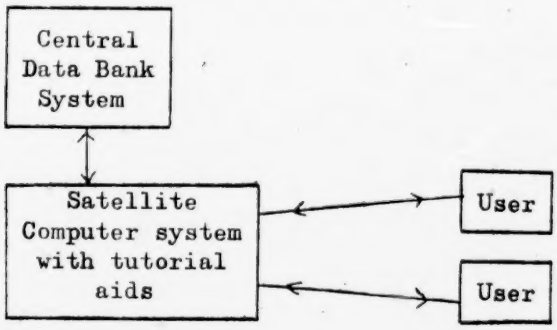


Figure 1

The flexibility that this approach can give is best illustrated when the central data-bank system comprises several data-bases as shown in Figure 2 below:

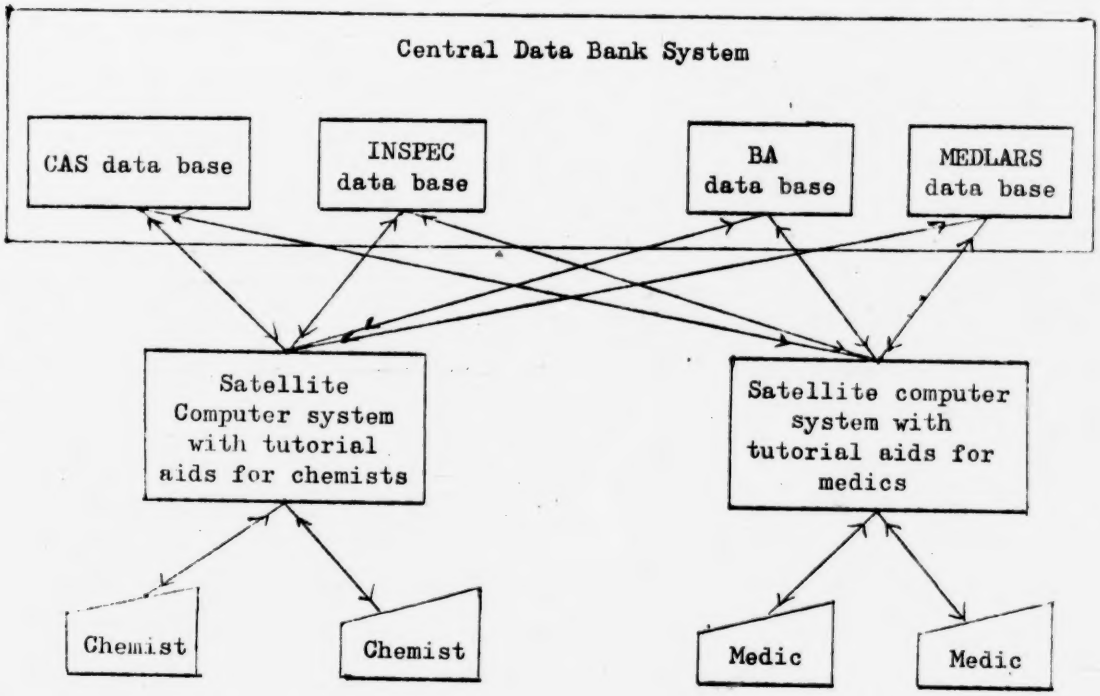


Figure 2

Each of the above satellite systems can be independently developed to enable a user - chemist or medic - to interface to any of the data-bases within the central data-bank using the language of his own subject area. Note that each satellite system will probably employ similar initial tutorial aids such as data-base description. Further advantages that can be realised by this approach are as follows:

a) Satellite systems can be developed as user needs dictate without affecting the operations of the central data-bank system and satellite already being

served by it.

b) Satellite systems, like the Newcastle University on-line Medlars system and Belfast's on-line Inspec system, have demonstrated that such systems can be developed and tested before the large central data-bank system is developed. Therefore there is no reason why development should not begin on satellite systems for other data-bases/user groups.

Problem Areas

A problem facing any designer of an on-line system is that of the availability of telecommunications software. For users of the current range of ICL computers this software is primitive. To develop an effective on-line system with such computers will require considerable resources to bring the basic telecommunications software up to an operational level.

There are two major reasons why existing manpower resources should not be utilised in this manner and they are:

a) ICL are in the process of producing a replacement for the current range of computers and this may render any telecommunications software developed obsolete on the new range of machines.
b) Telecommunications software development is an area of systems design and programming which should be viewed by us in much the same way as we view the development of compilers and operating systems.

It would be more fruitful to employ whatever manpower resources there are available in developing those aspects of on-line information retrieval which are germane to our current expertise and interest. This can only be achieved if such development work is carried out on hardware which has adequate telecommunications software support. Adopting this policy, we can still retain the initiative that systems such as the Newcastle on-line Medlars system has given us.

The other problem that we are faced with in this country is the lack of data communication facilities at a reasonable price. It would be possible to design a data network and predict what the cost of its use would be but as we are wholly dependent upon the GPO for communications facilities such a network must be many years off. In the meantime we perhaps have to accept that on-line searching is going to be expensive but if we are going to retain any expertise in this field we must continue with this work.

Appendix A3

*The BIRD On-line Retrieval System

L. D. Higgins

Department of Computer Science

At Belfast we have been developing an on-line reference retrieval system, BIRD. At present (September 1972) the BIRD data base allows retrieval from more than 5,000 recent papers taken from the physics abstracts section of the bi-monthly Inspec tapes on Atomic and Molecular Physics. A 'sceondary data base allows retrieval from the books in three local departmental libraries, namely numerical mathematics and computer science, physics,and applied mathematics - altogether about 1,000 books.

The talk proposes to outline the facilities BIRD offers to users, to discuss some of the problems that have arisen in implementing and maintaining BIRD, and to offer some guidelines of'user needs' as experienced by users of BIRD.

* Paper read at the Universities Computer Science Colloquium held in Edinburgh from September 19th to 22nd, 1972.

Introduction

The BIRD system is a full-text, on-line, mechanised, interactive system with multiple-term, subject search capabilities. The design of the system commenced five years ago and while it is presently being used in a local experimental operating environment, a dynamic research and development program continues to add new features and to increase overall system capacity. From the beginning consideration was given to cheaper disc storage space becoming available, to national teclecommunication networks being encouraged, and even in the longer term, to new technologies, such as laser storage techniques and microwave transmission.

Rather than provide a current awareness or retrospective service and so cater for static type profiles, we aim at offering a flexible service to the user who is prepared to build up his own profile interactively in his own time. The system in its present form can retrieve information from over 5000 abstract Inspec physics records and also from books in three departmental libraries, namely, 500 numerical and computer science books, 200 physics books, and 300 applied mathematics books. A picture of the overall system can be seen in the following illustration:

QUOBIRD SYSTEM OVERVIEW Slide 1

Circle (1) shows what data is selected for the primary Inspec data file and for the secondary book data file. I would like to draw your attention for a moment to another illustration which shows more clearly the subject classification headings that we extract from the bi-monthly Inspec tapes which we have been receiving for the past eighteen months:

INSPEC HEADINGS FOR PHYSICS CLASSIFICATION

Slide 2

Returning now to the former illustration:

QUOBIRD SYSTEM OVERVIEW

Slide 1

Circle (2) shows the actual data itself that is extracted from each book and from each Inspec tape record.

Circle (3) shows the machine readable format of the data before being transferred to disc. The information from the books is punched onto cards in an NIC/type format. I think it is worth mentioning here for the benefit of those familiar with NIC that we do not impose the constraints demanded by the NIC format. By that I mean that all the fields within each QUOBIRD unit record are variable lengths. This, of course, is more in keeping with the MARC type format in which the already established INSPEC data tape records are composed. I might add here that the reason for using NIC type format in the first place was that we thought at the beginning (when we started our whole project by indexing the books in the computer science departmental library) that we might want to use the NIC package system for producing catalogues and so forth. However, after closer consideration we did not think this worthwile.

Theoretical Aspects

Circle (5) implemented by circle (4) shows the file organisation that makes up the QUOBIRD files.

OVERALL FILE STRUCTURE Slide 3

This illustration shows the data file content and structure more clearly. It is a three-levelled inverted file processed by a hash indexing technique.

Back again to the first illustration:

QUOBIRD SYSTEM OVERVIEW Slide 1

Circles (6), (7) and (8) bring us back now to the on-line retrieval program, BIRD.

The facilities offered by BIRD can only truely be judged by a real-time on-line demonstration. However, since this is not on today's programme allow me instead to tell you a little more about BIRD and then I will show you a sample printout of what a novice user might experience during a typical on-line INSPEC data base search and what the experienced user might achieve ~~during~~ during the same search.

The BIRD system is:

Full-Text:

Every word (except those on the indexer-judged "noise word exclusion list") is indexed at the word level as a searchable term.

On-line:

The BIRD system operates in a real-time environment and uses the ordinary GPO telephone line as the communications link between user and system.

Interactive:

The full-text concept and on-line mode of operation permit a high degree of interaction between the user and the information with which he is working.

We have attempted to implement our present experimental pilot system under the following design criteria.

DESIGN CRITERIA

The user communicates with the system in ordinary English, and the dialogue guides him through each step of the research and retrieval process. For example, in response to the entry of a set of search terms, the system reports to the user the number of documents which satisfy his request and asks him whether he wishes to display the material or modify his search with additional terms.

The system's man-machine interface has been designed for simple yet effective operation. User training is minimal and experience by users has shown that one search session of half an hour is sufficient to be familiar with the mechanics of the system. We have, in addition, a small user manual which is easily read prior to a search or, if need be, referred to during a search. Other features of the system include a user help command facility, where a user can ask for guidance if he doesn't understand a prompt.

SAMPLE DISPLAY OF A NOVICE SEARCH

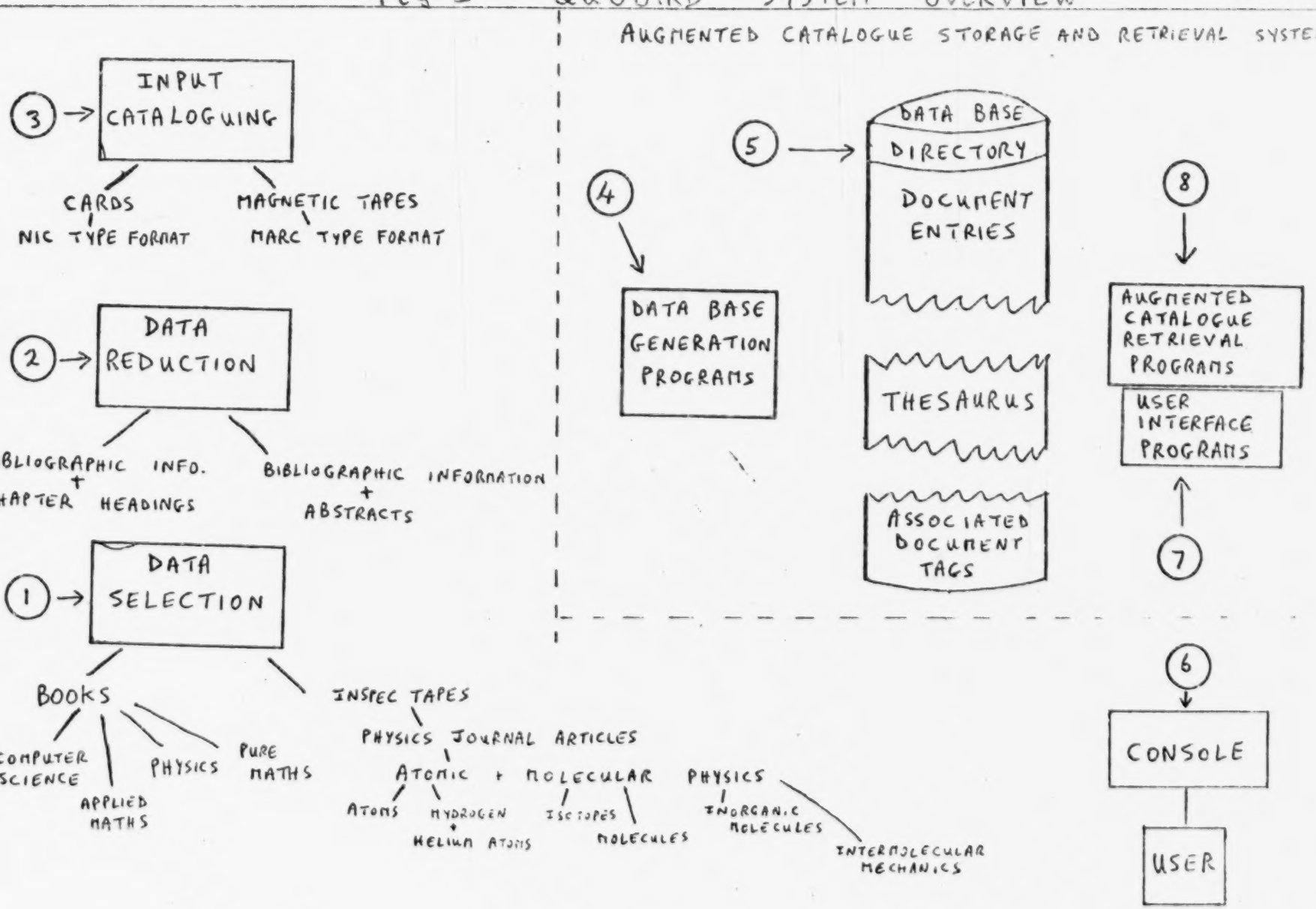
MULTI-TERM CLARIFICATION

SAMPLE DISPLAY OF AN EXPERIENCED SEARCHER

CURRENT PROBLEMS AND FUTURE USER NEEDS

SUCCESS OR FAILURE OF AN IR SYSTEM

Fig I QUOBIRD SYSTEM OVERVIEW

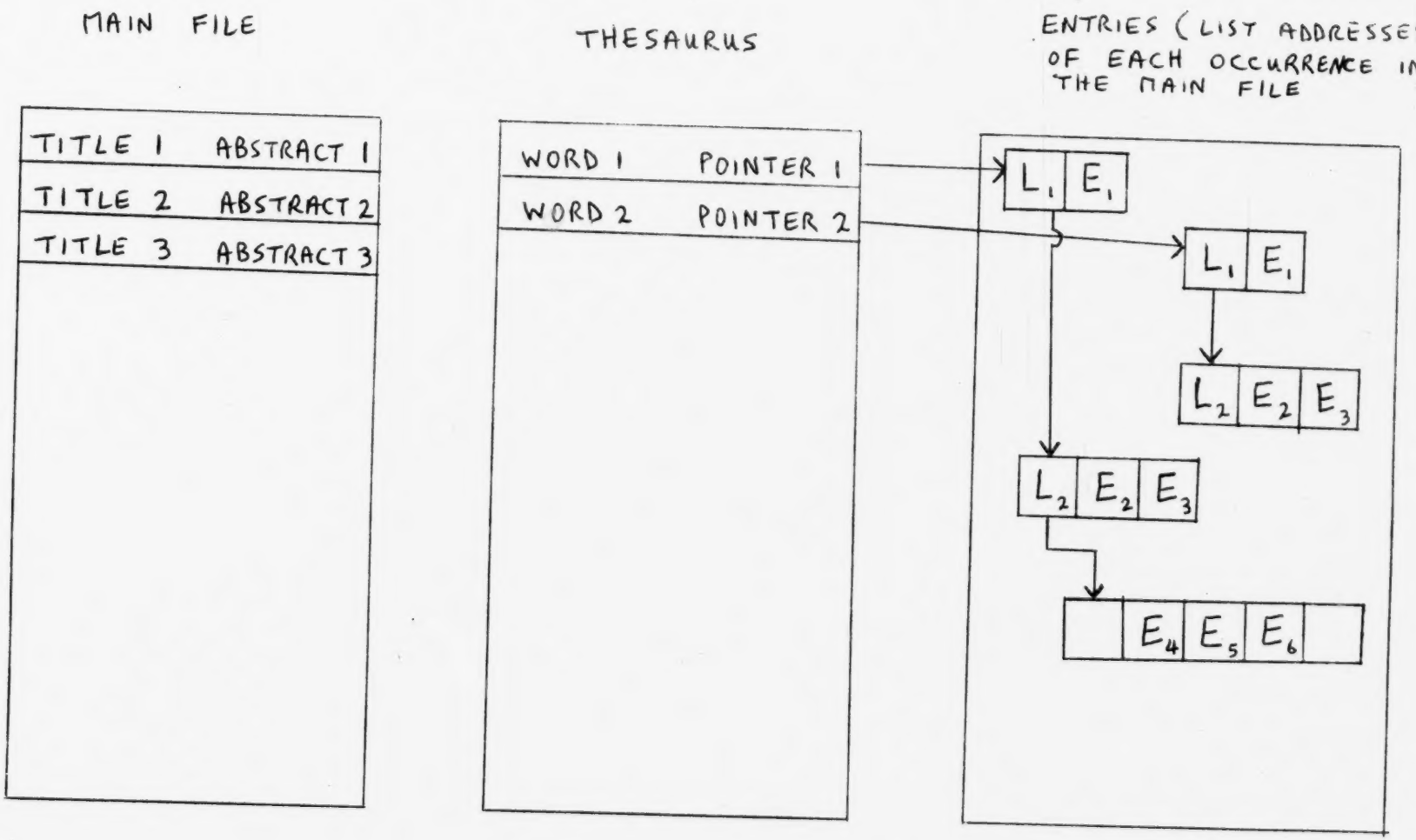


Slide 2

INSPEC HEADINGS FOR PHYSICS CLASSIFICATION

13.00	ATOMIC AND MOLECULAR PHYSICS
13.20	ATOMS
13.23	HYDROGEN AND HELIUM ATOMS
13.25	ISOTOPES
13.30	MOLECULES
13.31	INORGANIC MOLECULES
13.37	INTERMOLECULAR MECHANICS

OVERALL FILE STRUCTURE



Slide 4

DESIGN CRITERIA

(A) SELF INSTRUCTIVE

(B) SIMPLE TO USE

(C) RAPID RESPONSE

(D) EFFECTIVE TO USE

(E) MINIMUM COST

(F) EFFICIENCY OF STORAGE

(G) MINIMISING DISC ACCESSES

SC106 5

Sample Display of a Novice Search

```
BIRD:  INTERACTIVE SUBJECT INDEX OF ATOMIC AND MOL. PHYSICS RECORDS
BIRD:  ARE YOU FAMILIAR WITH THE SYSTEM?
    :  NO
BIRD:  TYPE "?" IF YOU NEED MORE INFORMATION AT ANY STAGE.
BIRD:  TYPE "X" IF YOU GIVE UP AND WISH TO STOP THE PROGRAM AT ANY TIME.
BIRD:  TYPE "A" IF YOU WANT ANOTHER SEARCH AT ANY TIME.
BIRD:  KEY?
    :  ?
BIRD:  TYPE A PHRASE,NO MORE THAN 8 WORDS.THIS (KEY) WILL BE USED AS A
BIRD:  UNIT FOR COMPARISON IN A SEARCH OF TITLES AND SUB HEADINGS
BIRD:   KEY?
    :  ARGON
BIRD:  KEY:  223 DOCS / 347 HDS
BIRD:  R,S,AB OR M?
    :  ?
BIRD:  R:  LIST OF REFERENCES REQUIRED
BIRD:  S:  HEADINGS THAT CONTAIN THE KEYS
BIRD:  AB: FULL TEXT OF ABSTRACT
BIRD:  M:  MORE KEY WORDS TO BE INCLUDED TO LIMIT OR INCREASE THE
           NUMBER OF DOCUMENTS RETRIEVED
BIRD:  R,S,AB OR M?
    :  M
BIRD:  KEY?
    :  EMISSION LINES
BIRD:  KEY:  60 DOCS / 78 HDS
BIRD:  INTERSECTION:  4 DOCS / 22 HDS
BIRD:  UNION:  279 DOCS / 419 HDS
BIRD:  N,I,P OR U?
    :  ?
BIRD:  N:  THE DOCUMENTS FOR THE NEW KEYS
BIRD:  I:  THE DOCUMENTS COMMON TO THESE AND PREVIOUS KEYS
BIRD:  P:  THE DOCUMENTS FOR THE PREVIOUS KEYS
BIRD:  U:  THE DOCUMENTS FOR BOTH THE NEW AND PREVIOUS KEYS
BIRD:  N,I,P OR U?
    :  I
BIRD:  R,S,AB OR M?
    :  X
BIRD:  THANK YOU AND GOOD DAY
BIRD:  DELETED:- OK
    :  £ENDJOB
MCS :  MCS (CORE:  640)
MCS :  CONNECT TIME  6:06   MILL TIME  0:933    DISC TRANSFERS  51
MCS :  LOGOUT LINE   2    MCPF    AEGD1254    12/37/54    15/09/72
```

Slide 6

Multi-Term Clarification

Simplest and most precise search involves a query about a subject which can be specifically described in a phrase, e.g. dissociative electron attachment in carbon dioxide.

A constraint on the number of documents recalled is imposed by using keywords in a phrase which must occur within a sentence to be recorded as a hit, e.g.

BIRD: KEY?

: RESONANCE SCATTERING

BIRD: KEY: 16 DOCS / 19 HDS

BIRD: KEY?

: RESONANCE

BIRD: KEY: 198 DOCS / 301 HDS

BIRD: R, H, T OR M?

: M

BIRD: KEY?

: SCATTERING

BIRD: NEW KEY: 190 DOCS / 369 HDS

: INTERSECTION: 30 DOCS / 77 HDS

: UNION: 358 DOCS / 650 HDS

Sample Display of an Experienced Searcher

```
BIRD:  INTERACTIVE SUBJECT INDEX OF ATOMIC AND MOL. PHYSICS RECORDS
BIRD:  ARE YOU FAMILIAR WITH THE SYSTEM?
    :  YES
BIRD:  KEY?
    :  INERT GASES
BIRD:  KEY:  37 DOCS / 52 HDS
BIRD:  R,S,AB OR M?
    :  M
BIRD:  KEY?
    :  ARGON
BIRD:  NEW KEY:  223 DOCS / 347 HDS
BIRD:  INTERSECTION:  5 DOCS / 12 HDS
BIRD:  UNION:  255 DOCS / 397 HDS
BIRD:  N,I,P OR U?
    :  U
BIRD:  R,S,AB OR M?
    :  M
BIRD:  KEY?
    :  ISOELECTRONIC EMISSION LINES
BIRD:  KEY:  3 DOCS / 4 HDS
BIRD:  INTERSECTION:  1 DOCS / 4 HDS
BIRD:  UNION:  257 DOCS / 400 HDS
BIRD:  N,I,P OR U?
    :  I
BIRD:  R,S,AB OR M?
    :  S
BIRD:  TITLE:
BIRD:  NEW OBSERVATIONS OF THE SPECTRA OF ARGON X TO XV AND OF ISOELECT
BIRD:  RONIC EMISSION LINES IN SILICON VII TO X, PHOSPHORUS X, SULPHUR
BIRD:  IX TO XII AND CHLORINE X TO XIV
BIRD:  THIS PAPER REPORTS THE CLASSIFICATION OF SPECTRAL LINES OF CHLOR
BIRD:  INE IX TO XIV AND OF ARGON X TO XV EMITTED FROM THE PLASMA FORME
BIRD:  D IN A THETA PINCH
BIRD:  THE WAVELENGTHS OF THE 2S/SUP 2/2P/SUP N/-2S2P/SUP N+1/ EMISSION
BIRD:   LINES ENABLE THE CALCULATION OF GROUND TERM INTERVALS IN THE SO
BIRD:  LAR ABUNDANT ELEMENTS SILICON, SULPHUR AND ARGON
BIRD:  THE MEASURED INTERVAL IN ARGON XIV ADDS CONFIRMATION TO THE IDEN
MCS :  BREAK:  6675  ON LINE LIMIT
    :  £RESUME
BIRD:  TIFICATION OF THE CORONAL FORBIDDEN LINE AT 4412 AA
BIRD:  USEFUL?
    :  YES
BIRD:  R,AB OR M?
    :  R
BIRD:  "NEW OBSERVATIONS OF THE SPECTRA OF ARGON X TO XV AND OF ISOELEC
BIRD:  RONIC EMISSION LINES IN SILICON VII TO X, PHOSPHORUS X, SULPHUR
BIRD:  IX TO XII AND CHLORINE X TO XIV
BIRD:   BY FAWCETT, B.C.  GABRIEL, A.H.  PAGET, T.M.
BIRD:  REF.NO.=297750 PUBLISHED IN J. PHYS. B (GB)    VOL. 4 NO.7    JU
       LY 1971
BIRD:  S,AB OR M?
    :  X
BIRD:  THANK YOU AND GOOD DAY
```

Problems in Implementation and Maintenance include:

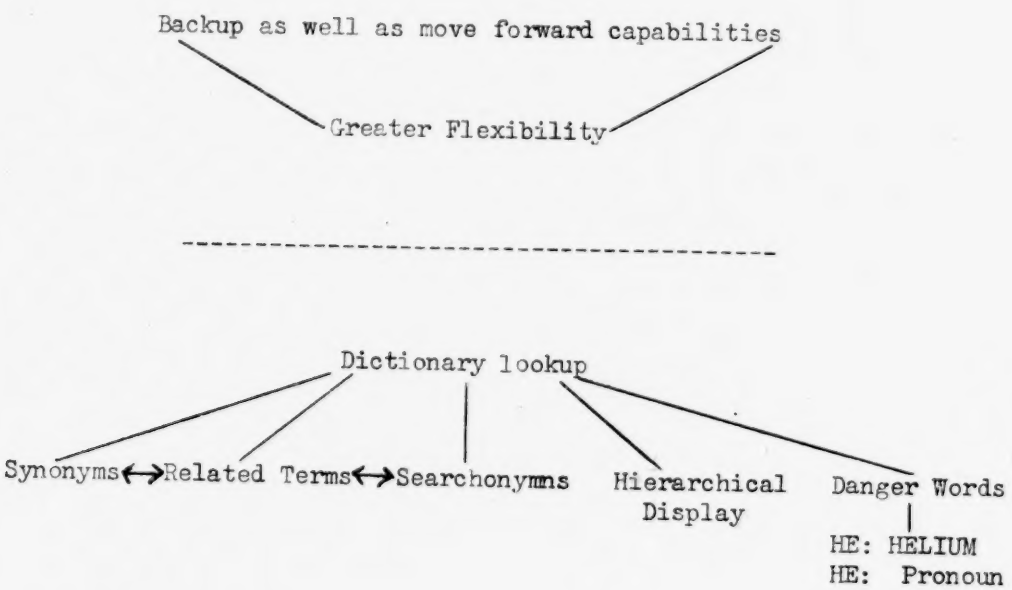
(a) System Security Requirements

(b) Text Editing Facilities Necessary

(c) System Expansion Requirements

(d) Hardware Failures

Guidelines for Future User Needs:



REFERENCES

1. L. D. Higgins and F. J. Smith: On-Line Subject Indexing and Retrieval, Program 1969,3, pp. 147-56.

2. L. D. Higgins and F. J. Smith: Disc Access Algorithms, Computer Journal, Vol.14, No.3, pp. 249 - 53, 1971.

3. M. Carville, L. D. Higgins and F. J. Smith: Interactive Reference Retrieval in Large Files, Inform. Stor. Retr. Vol.7, pp.205 - 210,1971.

4. Francis J. Smith: On-line Data Bank in Atomic and Molecular Physics, Physics of Electronic and Atomic Collisions, VII ICPEAC, 1971 - North Holland (1972).

Acknowledgement - This work is supported by the Office of Scientific and Technical Information, LONDON.

The success or failure of a computerised on-line IR system depends on:

1. Hardware/OS software backup
2. The indexing language adopted
3. The indexing techniques involved
4. Good User Interface Needs
5. System Evaluation

Appendix A4

Report on the Universities Computer Science Colloquium held in Edinburgh from September 19th to 22nd, 1972.

Although the colloquium lasted only two and a half days there was a heavy program covering a wide range of topics. These included computer education, computer simulation, information retrieval, programming languages and other miscellaneous topics. Probably the most interesting parts of the programme were the invited lectures; one by Professor Wirth on "Structured Programming" and another by Professor M. Wilkes on "The Hardware/Software Interface". Unfortunately another invited lecture on "Developments in the Theory of Computation" to be given by Dr. P. Landin had to be cancelled.

For those of us from Belfast one of the highlights of the programme was, of course, the talk on "The BIRD on-line reference retrieval system" given by Larry Higgins. This was very well received indeed. In addition, thanks to the Regional Computing Centre in Edinburgh, we had the unexpected opportunity to put on a live demonstration of the BIRD system in operation. This created a lot of interest and quite a number of the delegates were given the opportunity of operating the system themselves.

J. Boyle
29th September 1972

Appendix A5

REPORT ON 1972 CODATA CONFERENCE

1. Broadened Scope

One of the most significant aspects of the 1972 conference was the attention devoted to data in fields which have recently come within the scope of CODATA viz. Earth and Atmospheric sciences, Biological science, Astronomy and Astrophysics, and Engineering. The inclusion of these new fields was not universally welcomed and the point was voiced that the extension of CODATA activities, and in particular its extension to non-quantative data, could result in a dilution of effort which would have a detrimental effect in the subject fields originally covered by CODATA.

2. Data Evaluation

As in the 1970 conference there was some discussion of the problem of data evaluation although not as much. Possibly this was because it was felt that some progress was being made or, in some cases, because of the realisation that in some fields it is not possible, nor perhaps even desirable, to divert the major effort which would be required into data evaluation.

A number of talks were devoted to the problem of setting minimum standards of publication which could be used by editors, referees and authors to ensure that the reader is informed of the reliability of data published in a given article. Benson (session II) described the progress being made in the field of Chemical Kinetics while Westrum (session II) suggested a three levelled structure of publication rules. On the highest level, applying to all fields, would be the basic rules or "ten commandments" for publishing data. At an intermediate level there would be rules applying to specific fields while at the lowest level there would be detailed rules applying to particular areas within a field. A CODATA task group is working on the basic set of rules. The conference edition of CODATA NEWSLETTER (number 8) contained a copy of a paper called "A guide to procedures for the publication of Thermodynamic data".

3. Data Collections

The question was put as to whether in fact the effort devoted to the collection of data is always directed as effectively as it might. Is all the

data collected in fact useful? Are there areas to which more effort should be devoted? Engineers in particular seemed to feel that they are badly served by existing data collections. Wilkins(session V) said that in the field of Astronomy they were considering setting up a data centre as an alternative to a data bank. The function of the data centre would be to direct users to where the data they were seeking could be found. From the use made of the data centre it would be possible to determine whether in fact a data bank would be worthwhile. Schaefer(session VII) discussed the concept of an Information Analysis Centre whose function is to compress and evaluate data as a balance to the dilution and pollution which occurs in the normal course of events. The point was also made that the function of a data collection was not simply to supply the user with data but also to highlight "holes" in known data and to indicate potentially worthwhile areas of research.

6. Data and Computers

Criticism was expressed of the tendency for people to jump on the computer bandwagon and set up computerised data banks in cases where other forms of data collections, such as books, might fulfil the needs of the scientific community more efficiently. The point was made that the computer was most effective as a means of storing data when, in addition to storing and retrieving the data, the manipulative powers of the computer were used to process the data. Hilsenrath (session VI) suggested that many quantities which were previously stored in tabular form could now be most effectively stored by storing basic constants and the software necessary to calculate the required quantities from the basic constants.

Black (session II) stated that CODATA was setting up a roster or panel of experts in the field of computerised data storage from whom advice could be asked for by anyone considering setting up a computerized data bank. A symposium is being held next year by CODATA for experts in the field of computerised data storage.

J. F. Boyle
July, 1972

Appendix A6

ON-LINE 72 CONFERENCE

at Brunel University, 4 - 7th September, 1972

To accommodate the large audience a tent was erected but many of the slides presented were much too small for the size of the arena so that many of the lectures were difficult to follow.

There were generally four parallel sessions covering a wide variety of topics. Many of these were of a specialised nature e.g. "The use of computers in architectural design" or "The building of operating systems to control user programs working in an on-line mode".

Some lectures were given on information retrieval systems, the most lucid being one used by the U.S. Navy to monitor the state and position of their fleets throughout the world. This system used specially designed light pen displays on which maps of any part of the world can be flashed along with the data retrieved. This is a very complex system on which no expense is spared and systems such as these will be confined to military establishments for many years to come.

A group from Helsinki gave an interesting talk on editing text for newspapers. The computer hyphenated words and did some editing operations automatically but misspelling required user intervention.

An exhibition of peripheral equipment was given. One of the most interesting exhibits was a teletypewriter which had a cassette tape unit. Information could be stored and edited on the tape and then, when error free, transmitted to the computer at high speed.

H. O'Hara

Appendix A7

QUEEN'S UNIVERSITY INFORMATION SYSTEMS

Dear Sir,

References on Atomic and Molecular Physics

An information system for the retrieval of references on Atomic and Molecular Physics is now available to all members of the University through any teletype terminal or Visual Display Unit. A VDU is now in the School of Applied Mathematics and Physics Library.

At present we have nearly 7,000 references in the system, consisting of papers published over the last two or three years and we are up-to-date with Physics Abstracts. Each reference consists of an abstract and the usual bibliographic details of author, journal title, volume, number, etc.

You are now invited to use the system. We think it will be useful to you and it should allow you to retrieve papers after a search in greater depth than is possible manually with Physics Abstracts. We would appreciate your comments - good or bad. Only through practical use of the system can we hope to assess its facilities.

Instructions for logging in to the system are enclosed. Our recommendation is that you use the system briefly and then request the user manual which gives further guidance.

Books in the School Library

A secondary system allows the retrieval of books from the School Library. Titles and chapter headings may be searched for subjects on which information is required.

For further details please refer to:- Mrs. Joan Stewart
extension 489 (G.P.O.).

Yours faithfully,

L. D. Higgins
Projects Manager

USING A VISUAL DISPLAY UNIT

FOR ATOMIC AND MOLECULAR PHYSICS ABSTRACTS

Type : ØJOB,PUBL,ABED1234
Press ESCAPE
MCS : PASSWORD?
Type : 8 spaces
Press ESCAPE
MCS : LOGIN LINE ----------
Type : ØRUN,ICL,ABST,RETR
Press ESCAPE

FOR APPLIED MATHEMATICS AND PHYSICS BOOKS

Type : ØJOB,PUBL,ABED1234
Press ESCAPE
MCS : PASSWORD?
Type : 8 spaces
Press ESCAPE
MCS : LOGIN LINE ----------
Type : ØRUN,ICL,AMPB,RETR
Press ESCAPE

FOR COMPUTER SCIENCE BOOKS

Type : ØJOB,PUBL,ABED1234
Press ESCAPE
MCS : PASSWORD?
Type : 8 spaces
Press ESCAPE
MCS : LOGIN LINE ----------
Type : ØRUN,ICL,CSCB,RETR
Press ESCAPE

User actions are given on the left and are underlined. "Press" refers to depressing the "ESCAPE" button after each instruction has been typed.

Type : ØRESUME after a "BREAK" message from MCS
Type : ØEND JOB after "X" has been used to delete the program.

U S I N G A T E L E T Y P E

FOR ATOMIC AND MOLECULAR PHYSICS ABSTRACTS

Type	:	£JOB,PUBL,ABED1234
Press		ACCEPT
MCS	:	PASSWORD?
Type	:	8 spaces
Press		ACCEPT
MCS	:	LOGIN LINE ----------
Type	:	£RUN,ICL,ABST,RETR
Press		ACCEPT

FOR APPLIED MATHEMATICS AND PHYSICS BOOKS

Type	:	£JOB,PUBL,ABED1234
Press		ACCEPT
MCS	:	PASSWORD?
Type	:	8 spaces
Press		ACCEPT
MCS	:	LOGIN LINE ----------
Type	:	£RUN,ICL,AMPB,RETR
Press		ACCEPT

FOR COMPUTER SCIENCE BOOKS

Type	:	£JOB,PUBL,ABED1234
Press		ACCEPT
MCS	:	PASSWORD?
Type	:	8 spaces
Press		ACCEPT
MCS	:	LOGIN LINE ----------
Type	:	£RUN,ICL,CSCB,RETR
Press		ACCEPT

User actions are given on the left and are underlined. "Press" refers to depressing the "ACCEPT" button after each instruction has been typed.

Type : £RESUME after a "BREAK" message from MCS

Type : £ENDJOB after "X" has been used to delete the program.

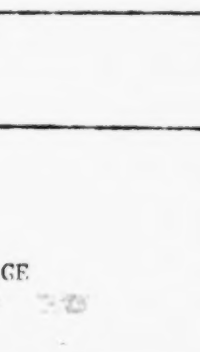
To: Console Representative

From: Computer Advisory Service

Please display the enclosed posters adjacent to your Teletypewriter. They should be attached in sequence as shown below:-

OPERATING THE CONSOLE

THE CONSOLE



STARTING

LOGGING IN

LOGGING OUT

CANCELLING A MESSAGE

CONSOLE MESSAGES

STOPPING

FAULTS

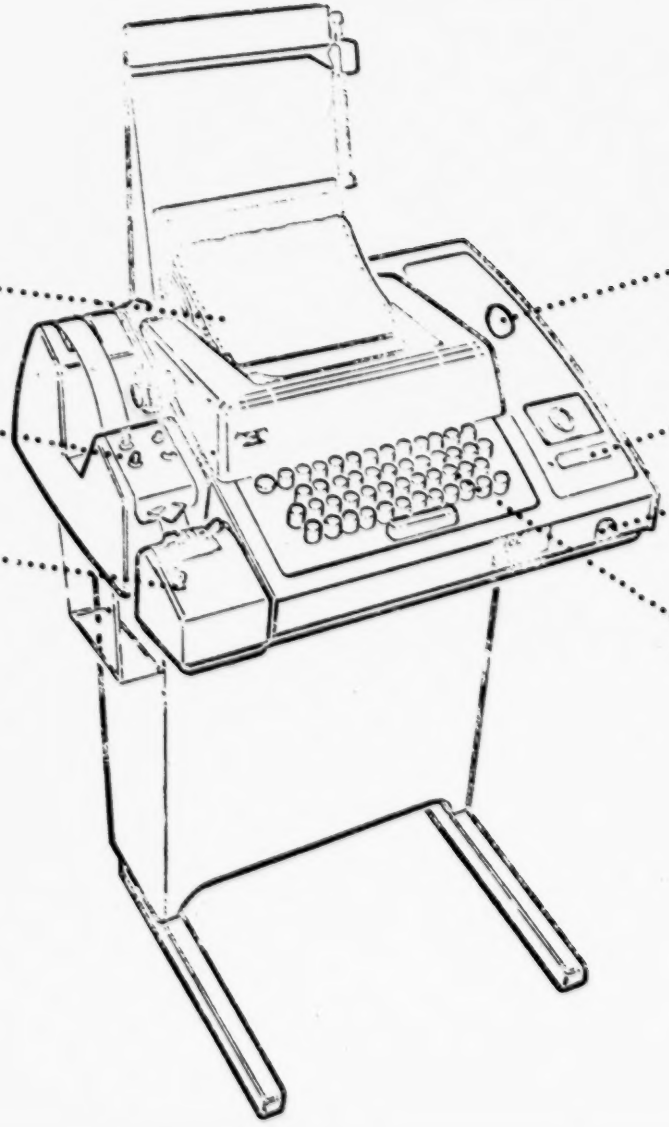
Thank you,

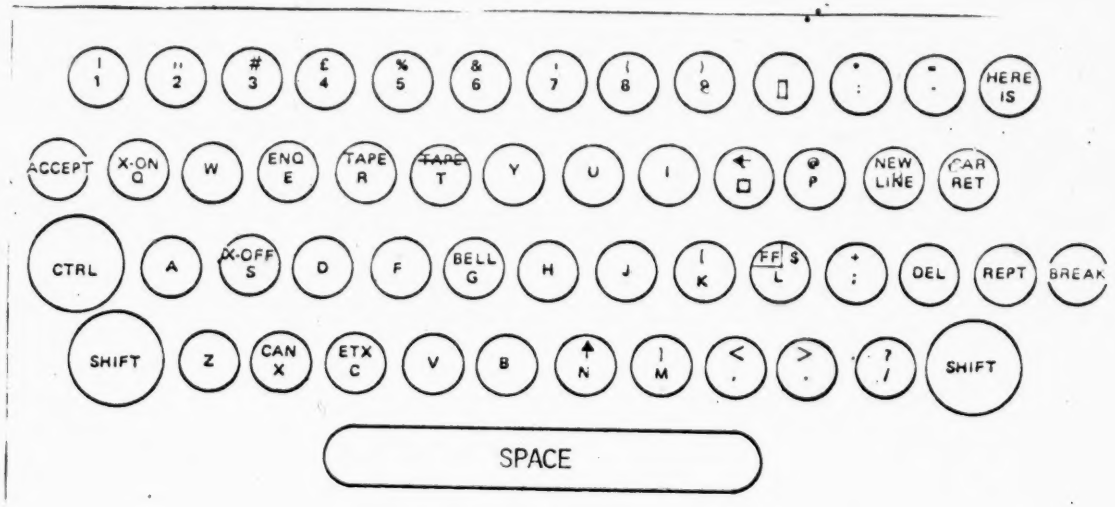
J. Kingman S. Gregg

R. GREGG
Jan. 73.

~ OPERATING THE CONSOLE ~

THE CONSOLE





STARTING

1. CHECK PAPER IN CONSOLE
2. SWITCH ON MAINS SUPPLY
3. SWITCH FUNCTION SWITCH TO **ON-LINE**
4. PRESS **START** BUTTON

   IF THE TYPEWRITER 'CHATTERS', RING COMPUTER RECEPTION AND REQUEST YOUR CONSOLE BE CONNECTED TO THE COMPUTER
5. PRESS **CTRL** AND **A** TOGETHER AND **MCS** WILL TYPE

   MCS : READY (VERSION

CONSOLE MESSAGES

/ TIMED OUT

THIS OCCURS AFTER 60 SECONDS OF NON-ACTIVITY ON THE CONSOLE. TO PROCEED PRESS

CTRL AND **A** TOGETHER

/ LOST INPUT

THIS MESSAGE WILL OCCUR IF THE CONSOLE FAILS TO TRANSMIT THE INPUT CORRECTLY.

RE-TYPE THE MESSAGE

LIMIT ON LINES AND TIME

WHEN A LINE OR TIME LIMIT IS REACHED MCS SUSPENDS ACTION WITH THE MESSAGE - MCS : BREAK LOCATION ON $\frac{\text{LINE}}{\text{TIME}}$ LIMIT

TO CONTINUE TYPE **£ RESUME**

STOPPING

1. ENSURE YOU HAVE **LOGGED OUT**
2. SWITCH FUNCTION SWITCH TO **OFF** OR **LOCAL**
3. SWITCH **OFF** MAINS SUPPLY
4. REPLACE COVER OVER CONSOLE

FAULTS

IF YOU FIND OR SUSPECT A CONSOLE FAULT, PLEASE COMPLETE A **CONSOLE FAULTS REGISTER** FORM AND SEND SAME TO COMPUTER ADVISORY SERVICE.

LOGGING IN

TO A PUBLIC FILE

£JOB, PUBL, JOB NUMBER

TO A PRIVATE FILE

£JOB, FILENAME, JOB NUMBER

LOGGING OUT

TYPE **£ENDJOB**

IT IS IMPERATIVE THAT A USER LOGS OUT BEFORE STOPPING THE CONSOLE

CANCELLING A MESSAGE

TO PREVENT A SINGLE CHARACTER FROM BEING TRANSMITTED TO THE PROGRAM PRESS **DEL** OR **RUB OUT**

TO PREVENT A COMPLETE LINE OF CHARACTERS BEING TRANSMITTED TO THE PROGRAM PRESS **CTRL** AND **X**

Appendix A8

Jahreskolloquium zur Rechentechnik

Dienstag, 29. Februar 1972

Technische Universitaet Braunschweig
Pockelsstrasse 4 (Hauptgebaeude) Hoersaal S4

Kurzfassungen der Vortraege

F. J. Smith: "R.I.O.T.: Retrieval of Information On-line by Telephone".

We have been building two on-line information systems here in Belfast: one for the retrieval of references and the other for the retrieval of numerical atomic data. They are designed for retrieval via a computer terminal linked to our data bank through the ordinary telephone network; so anyone who can dial the Belfast exchange can connect with our system, and through a simple question and answer form of conversation with the system can, we hope, retrieve the information he needs. Much emphasis has been placed in our work on the efficiency of the software, as the main stumbling block to the development of this kind of information service is the great cost. By careful study of the techniques used we think we have reduced these costs by a considerable factor.